Semi-parametric Regression Models for Risk Differences, Rate Differences and Relative Risks

By

Mark William Donoghoe

A thesis submitted to Macquarie University

for the degree of

Doctor of Philosophy

Department of Statistics

Faculty of Science and Engineering

June 2015

MACQUARIE University

SYDNEY·AUSTRALIA

Contents

Appendices

Abstract

Two fundamental biostatistical measures are the risk and the rate of event occurrence, representing the probability of an event and the expected number of events during a fixed time period. Regression models can be used to relate an individual's characteristics to the risk or rate of an event, such as the occurrence of disease or death. This allows identification of high-risk individuals and can reveal ways in which risk may be reduced.

Generalised linear models (GLMs) for binary and count data are an important statistical tool for risk and rate modelling, and semi-parametric extensions provide additional flexibility. However, some key GLMs of interest have parameter constraints implied by the risk and rate models, and standard model-fitting algorithms can be numerically unstable. This is particularly true for GLMs that allow estimation of risk differences, rate differences and relative risks.

In this thesis by publication, new variants of the Expectation–Maximisation (EM) algorithm are developed in order to provide reliable and flexible methods for fitting such models to binary and count data. This begins with the development of a method for additive binomial GLMs, which allows for reliable adjustment of risk differences. An extension of this and other EM-type algorithms for binomial and Poisson GLMs is then provided, which allows for flexible semi-parametric regression based on spline models. As well as risk differences, these models allow reliable estimation of rate differences and relative risks. A method for additive regression under a negative binomial model is also developed, which can be used to estimate rate differences when the observed counts show more variation than is expected under a Poisson model. These methods all ensure that the fitted models respect the required parameter constraints, and their stability

allows us to reliably use resampling methods that require many auxiliary analyses, such as the bootstrap.

The utility of these approaches is demonstrated by applying them to various clinical datasets. The methods described in this thesis have all been implemented in open-source packages for the R computing environment and have been made available online.

Statement of candidate

I certify that the work in this thesis entitled "Semi-parametric Regression Models for Risk Differences, Rate Differences and Relative Risks" has not previously been submitted for a degree, nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any assistance that I have received in my research work and the preparation of the thesis itself has been appropriately acknowledged.

In addition, I certify that all information sources and literature used are referenced in the thesis.

Mark William Donoghoe

Date

Acknowledgements

First and foremost, I would like to thank my principal supervisor, Professor Ian Marschner, for giving me the opportunity to undertake this PhD, and for his patience, guidance and knowledge that helped me to see it through. I can't overstate my gratitude.

I'd also like to express my appreciation to my associate supervisor, Associate Professor Gillian Heller, and the other staff and post-graduate students in the Department of Statistics at Macquarie University for providing a good atmosphere for research, and to my colleagues at the NHMRC Clinical Trials Centre for their ongoing support.

My research was supported by a Macquarie University Research Excellence Scholarship (MQRES), associated with an Australian Research Council grant (DP110101254). I gratefully acknowledge the generosity of the ASSENT-2 and ASSENT-3 investigators, Professor Tony Keech (FIELD study) and Professor Brian Darlow (BOOST-NZ study) for kindly allowing me to analyse and present data from their studies in this thesis.

A special thanks to my wonderful family for helping me get to this point, and to my friends and football for keeping me in good spirits throughout.

And last but not least, to Divya. Your unwavering love has been invaluable, and I hope I can somehow repay the sacrifices that you made to make this possible for me.

Publications and presentations

Peer-reviewed journal articles

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* **80**: 184–196. DOI: 10.1016/j.csda.2014.06.019.

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* **11**(1): 91–108. DOI: 10.1515/ijb-2014-0044.

Manuscript ready for submission

Donoghoe, M. W. and I. C. Marschner. Additive negative binomial regression for overdispersed count data with application to adjusted rate differences. Manuscript ready for submission.

Software packages

Donoghoe, M. W. (2015a). *addreg: Additive Regression for Discrete Data*. R package version 2.0. URL: http://CRAN.R-project.org/package=addreg.

Donoghoe, M. W. (2015b). *logbin: Relative Risk Regression Using the Log-Binomial Model*. R package version 1.2. URL: http://CRAN.R-project.org/package=logbin.

Conference presentations

Donoghoe, M. W. *Adjusted risk difference estimation using a stable and flexible method for additive binomial models.* Presented at the Australian Statistical Conference / IMS Annual Meeting, Australian Technology Park, Sydney, July 7–10, 2014. Awarded SSAI Certificate of Merit.

Donoghoe, M. W. *Smooth semi-parametric adjustment of rate differences, risk differences and relative risks.* Presented at the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, July 14–18, 2014.

Conference proceedings paper

Donoghoe, M. W. and I. C. Marschner. Smooth semi-parametric adjustment of rate differences, risk differences and relative risks. *Proceedings of the 29th International Workshop on Statistical Modelling.* Ed. by T. Kneib, F. Sobotka, J. Fahrenholz, and H. Irmer, 2014. **1**: 105–110.

List of Figures

List of Tables

1

Introduction

An important part of biostatistics is the study of factors that affect the occurrence of events relevant to human health. In this thesis we will develop new regression methodology motivated by biostatistical applications, focusing on the relationship between risk factors and health events.

The most important event relevant to the health of an individual is death. Accordingly, mortality is a common outcome in biostatistical applications, but there are nonetheless many other types of events that are significant in the study of human health. Commonly, such events are the manifestation of an underlying disease process, and may be the trigger for a patient to be hospitalised, to undergo a corrective or preventative procedure, or to change their ongoing treatment approach. These events and their subsequent consequences therefore usually have a negative impact on health status and quality of life, and increase the financial burden on the individual and the public health system. Therefore, our motivation for studying such events is typically to reduce the number of events that occur on both an individual and population level. On the other

hand, an event of interest may instead involve a positive outcome, such as a cessation of disease symptoms, in which case the goal will be to increase its frequency. In either case, an essential step towards achieving the desired objective is to improve our understanding of the factors that contribute to the occurrence of the events of interest. Statistical models of event occurrence require a quantitative measure of such occurrence. In this thesis we will develop methodology for models of event risks and event rates, which are two of the most fundamental biostatistical measures. We therefore begin with a brief discussion of the concepts of risk and rate, before moving onto a discussion of how these measures will be studied in this thesis.

1.1 Risk and rate modelling

In biostatistics, an individual's *risk* of an event is the probability of the event occurring within a fixed period of time (Lachin, 2011). As such, risk is always defined with reference to a specific time period of interest that is common to all individuals. In this thesis, we will be particularly interested in modelling the relationship between risk and individual characteristics by using regression models. The risk for an individual with characteristics specified by the covariate vector $\boldsymbol{x}$ is

$$\text{Risk}(\boldsymbol{x}) = \Pr(\text{event} \mid \boldsymbol{x}). \tag{1.1}$$

As a measure of event susceptibility, risk is relevant in situations involving binary outcomes, that is, those that can take one of only two possible values. In this context, we denote these values as 'event' and 'no event', but the same framework can be applied in a wide variety of similar scenarios. In biostatistics, these outcomes are modelled as the realisation of a Bernoulli random variable, or more generally, a binomial random variable. The assumption underlying the binomial distribution is that the risk, or the probability of an event, is constant across a fixed number of independent trials, which would typically correspond to multiple individuals with the same characteristics observed in a single fixed time period. The number of events cannot exceed the number of trials, and this binomial model includes the special case of a single trial for each observation, such that only one event is possible within the risk period. Binomial regression, including binary regression, is therefore a key methodology in the modelling

of risk.

Alternatively, the outcome of interest may be such that multiple events are possible over time and there is no upper limit on the number of events that can occur. In this context, the event *rate* is defined as the expected number of events in one unit of time. Thus, the rate for an individual with characteristics $\boldsymbol{x}$ is

$$\text{Rate}(\boldsymbol{x}) = \mathbb{E}(\text{events in 1 time unit} \mid \boldsymbol{x}). \tag{1.2}$$

Unlike the definition of risk, event rates are defined per time unit and therefore allow for different periods of observation for different individuals.

Events over time are often modelled as the realisation of a Poisson process, so that the event count has a Poisson distribution with mean equal to the event rate per time unit multiplied by the observation time. Such count outcomes can also be modelled as the realisation of many binary trials, where the risk of an event within a small time period depends on the underlying event rate. Under some weak conditions, as the time period for each hypothetical binary trial becomes infinitesimally small, the resulting number of events will again have a Poisson distribution (Cameron and Trivedi, 1998, p. 5). Poisson regression is therefore an important methodology in the modelling of event rates associated with count data outcomes.

However, the Poisson distribution has the property that the variance of the number of events is equal to the expected number of events, which often does not hold true in real data. We commonly observe *overdispersion*, in which the variance exceeds the expected value, meaning that a Poisson model is inappropriate. An alternative distribution for modelling count data in the presence of overdispersion is the negative binomial distribution, which allows for heterogeneity between individuals beyond that captured by their observed characteristics (Greenwood and Yule, 1920). As with the Poisson model, the expected total number of events for a single observation is dependent on the fixed period of exposure and the average event rate, which may be affected by certain individual characteristics.

One purpose of regression modelling in the above contexts is to examine the relationship between individual characteristics and the risk or rate of events in the population. This may assist in understanding complex disease processes by highlighting factors that

form part of the direct causal chain, or act indirectly through their association with the direct causes. The resulting model could be used to identify high-risk individuals such that personalised preventative measures can be undertaken, as well as to highlight modifiable risk factors that could potentially be targeted as part of a new intervention. By quantifying the effects of these risk factors, we are able to estimate the impact that a successful approach would be expected to have on the number of events that occur, as well as on secondary outcomes such as public health costs.

Regression modelling is also useful in controlled clinical trials, in order to quantify the effect of the intervention under study while also taking into account other variables that impact on the risk or rate of events. In a randomised study, we expect that important risk factors will be evenly distributed between the treatment groups, but adjusting for these will account for differences that may have occurred due to chance, and will usually produce a more efficient estimate of the effect of interest (Hauck, Anderson, and Marcus, 1998). An accurate estimate of the effect of the intervention allows for an assessment of its potential utility, particularly in terms of the trade-off between the expected number of events that could be prevented and the cost of its introduction.

A further application of regression modelling is to identify interactions. A statistical interaction exists if the magnitude of the effect of one characteristic on the risk or rate differs depending on the value of a second characteristic (Rothman, Greenland, and Walker, 1980). Identification of interactions can further help to improve understanding of the underlying disease mechanisms, and in the case of a clinical trial, may identify subgroups of patients in which the intervention is expected to be more or less efficacious.

1.2 Effect measures

In quantifying the impact of a risk factor or an intervention on the risk or rate of an event, an appropriate effect measure must be used. This refers to the scale on which the change in risk or rate is expressed. A large number of studies have shown that the choice of effect measure can have a substantial impact on the interpretation of the effect itself, and hence influence subsequent treatment decisions (e.g. Forrow, Taylor, and Arnold, 1992; Bobbio, Demichelis, and Giustetto, 1994; Lacy et al., 2001). This suggests that the effect measure should be chosen based on interpretability.

Different biological models lead to different relationships between covariates and the risk or rate of events, and this should be taken into consideration when assessing potential interactions (Walter and Holford, 1978). But when the biological mechanism is unknown, Rothman (1978) argues that the best summary of the treatment effect is that which gives the most parsimonious model. This suggests that model fit relative to the number of parameters in the model should also be a factor in choosing an effect measure.

Risk and rate effect measures are often expressed on a relative scale, that is, as a multiplicative effect on the baseline risk or rate. However, if there is a constant absolute risk or rate difference across strata, this will appear as a statistical interaction on a relative scale (e.g. Marschner, Gillett, and O'Connell, 2012), and an additive model would provide a more appropriate summary of the data.

The choice of effect measure may also be driven by the aims of a particular study. Relative measures, provided they are presented alongside baseline risks and rates (Berry, Knapp, and Raynor, 2006), are useful for identifying factors that have a large effect on individual risk. On the other hand, absolute measures may be preferred from a public health perspective (Egger, Smith, and Phillips, 1997). In particular, an estimate of the effect of the intervention or risk factor on an absolute scale can be directly related to the change in the number of events that we would expect on a population level as a result of introducing the intervention or modifying the risk factor in question.

In the following subsections we review some of the principal effect measures used for binary and count data.

1.2.1 Odds ratio

The odds ratio is often used as the effect measure for a binary outcome. If $p_1 = \text{Risk}(\boldsymbol{x}_1)$ denotes the risk of an event for an individual with some characteristic $\boldsymbol{x}_1$, as specified in (1.1), and $p_2 = \text{Risk}(\boldsymbol{x}_2)$ the risk for an otherwise similar individual with characteristic $\boldsymbol{x}_2$, then

$$\text{Odds ratio } (\boldsymbol{x}_2 \text{ vs. } \boldsymbol{x}_1) = \frac{p_2/(1-p_2)}{p_1/(1-p_1)}.$$

An important property of the odds ratio is that it is symmetric with regards to the outcome definition. That is, if events and non-events are interchanged, then the odds

ratio is inverted (Walter, 2000). Furthermore, the odds ratio is the only estimable effect measure in a retrospective case-control study. Outside of case-control studies, odds ratios are widely used because they result naturally from logistic regression models, which have favourable mathematical properties. This is partially because no limit is required on the range of the logarithm of the odds ratio: an odds ratio anywhere in $[0, \infty)$ can be applied to any $p_1 \in [0, 1]$ to produce a valid risk $p_2 \in [0, 1]$. As a result, Cook (2002) argues that odds ratios may be relevant in a broader population of patients than the other common relative effect measure, the relative risk, which we discuss in the next subsection.

Although it has some theoretical advantages, the odds ratio has received criticism as an effect measure because its interpretation is not straightforward, and it is often reported as if it were a relative risk (Holcomb et al., 2001). In situations in which the event of interest is rare, the odds ratio is a close approximation to the relative risk, so the distinction between them may be inconsequential. However, the odds ratio will always be more extreme than the relative risk, and hence it may be seen to exaggerate the size of the effect (Davies, Crombie, and Tavakoli, 1998). Such exaggeration can be substantial when the outcome of interest is not rare, and so odds ratios are often viewed as an undesirable effect measure in prospective studies of common outcomes.

1.2.2 Relative risk and rate ratio

Relative risks and rate ratios provide a more easily interpretable effect measure on a multiplicative scale. As above, if p_1 and p_2 represent the risk or rate for individuals with characteristics $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, as specified in (1.1) and (1.2), then

$$\text{Relative risk } \textit{or} \text{ rate ratio } (\boldsymbol{x}_2 \text{ vs. } \boldsymbol{x}_1) = \frac{p_2}{p_1}.$$

Unlike the odds ratio, the relative risk is not symmetric with respect to the outcome definition. However, it does have another desirable property, called collapsibility. This means that, in the absence of confounding, the overall relative risk is equivalent to the weighted average of stratum-specific relative risks, and the unadjusted effect measure will not change after adjustment for a covariate that is not a confounder (Cummings, 2009). This is not true of the odds ratio, which can hence show apparent confounding

even where none exists.

In order to ensure that the resulting risk p_2 lies within $[0, 1]$, the permitted range of the relative risk is dependent on the baseline risk p_1: it must lie in $[0, 1/p_1]$. On the other hand, because rates have no upper limit, there is no similar limit on the rate ratio, which can take any non-negative value. The rate ratio is the natural effect measure that results from Poisson regression, and as a result, it is widely used in regression analysis of count data. Although the same is not true for the relative risk in binomial regression models of binary outcomes, it has been increasingly recommended in this context (Wacholder, 1986; Blizzard and Hosmer, 2006; Lumley, Kronmal, and Ma, 2006).

1.2.3 Risk and rate difference

The risk and rate difference are absolute effect measures, where

$$\text{Risk difference } \textit{or} \text{ rate difference } (\boldsymbol{x}_2 \text{ vs. } \boldsymbol{x}_1) = p_2 - p_1,$$

using the same notation as above. As opposed to the multiplicative effect measures, the risk difference and rate difference allow for calculation of the effect of a risk factor or intervention on the absolute number of events that are expected to occur.

In two groups of equal size n, with average risks of p_2 and p_1, we expect np_2 and np_1 events respectively. The difference in the expected number of events is therefore $np_2 - np_1 = n \times$ (Risk difference). If these groups are representative of a hypothetical population before and after the introduction of an effective intervention, the risk difference can be used in a cost-effectiveness analysis, to estimate the amount of money that could be saved by preventing events, in order to compare it to the cost of implementing the intervention (Willan and Briggs, 2006).

The risk difference is closely linked to the concept of the number needed to treat (NNT), which estimates the number of patients that would need to receive a novel intervention in order to avoid a single event during the fixed time period (Laupacis, Sackett, and Roberts, 1988). The NNT is sometimes cited as a more clinically meaningful representation of a treatment effect (Cook and Sackett, 1995), but some care must be taken with its interpretation (Kristiansen and Gyrd-Hansen, 2004). The risk difference is

both symmetric and collapsible, but given $p_1 \in [0, 1]$, it must lie in $[-p_1, 1 - p_1]$ to ensure that $p_2 \in [0, 1]$.

Similarly for rates, $n \times$ (Rate difference) is the expected difference in the number of events occurring in exposure time n for two groups with rates p_2 and p_1. Thus it can be used in cost-effectiveness analyses in a similar way to the risk difference, but caution must be exercised when the exposure time for an individual is affected by the occurrence of events, or if it is dependent on the individual's covariates (Greenland, 1996). The permissible range of the rate difference, given p_1, is $[-p_1, \infty)$, in order to ensure that $p_2 \geq 0$.

1.3 Semi-parametric regression

The standard methods for relating a continuous covariate to the risk or rate of an event require that the form of the relationship between the covariate and the outcome is specified in advance. A linear or log-linear relationship is often used, but others can be implemented by applying an appropriate transformation to the covariate in question. However, this requires that we have some prior information about the form of the relationship, which is often not available when we are investigating a new risk or rate model. Semi-parametric regression is a method that removes the constraint of a fully parametric model, allowing for additional flexibility in the relationship between a covariate and the outcome by assuming only that it belongs to a flexible class of functions.

Semi-parametric regression provides an estimate of the shape of this covariate relationship, which may allow a better representation of the underlying biological association than one in which the functional form has been precisely specified. This may allow us to identify a parametric functional form that better approximates the true relationship, producing a more parsimonious model and providing a more appropriate estimate of the effect size associated with a continuous covariate. For example, if the semi-parametric model suggests that the effect of the covariate on the risk or rate of an event takes a piecewise linear form, we can use this as the functional form in a fully parametric model in order to estimate the size of the effect below and above the change point.

Additionally, semi-parametric regression is useful in situations where the effect of a

continuous covariate is not of direct interest, but we wish to adjust for it in estimating the effect of an intervention or another risk factor. In particular, if we wish to estimate a particular effect measure for the characteristic of interest, we must use an appropriate model, as discussed in Section 2.1.1. A fully parametric model imposes constraints on the possible forms of the relationship between other covariates in the model and the risk or rate. With a semi-parametric model, however, we can estimate the effect of interest but relax the restrictions on the effect of other covariates. As a result, we can expect that the adjustment is a more realistic representation of reality, and the estimated effect of interest should be more accurate.

1.4 Example datasets

Throughout this thesis, examples from real datasets are used to demonstrate the utility of the methods described. These datasets, and the context of the studies that generated them, are summarised briefly below.

1.4.1 ASSENT-2 study

The Assessment of the Safety and Efficacy of a New Thrombolytic (ASSENT-2) study was a double-blind randomised trial to compare two fibrinolytic therapies in 16,949 patients with acute myocardial infarction (ASSENT-2 Investigators, 1999). This type of treatment aims to dissolve blood clots so that blood flow can be stabilised. The treatments being studied were the gold standard, alteplase, versus a genetically engineered variant, tenecteplase, which allows for easier administration.

A previous study estimated a dose of the new therapy with equivalent efficacy as alteplase (Van de Werf et al., 1999), and the aim of ASSENT-2 was to test this hypothesis. The primary endpoint of the study was all-cause mortality within 30 days, a binary outcome. A number of individual characteristics that might affect risk of death were collected at baseline, and the primary analysis included adjustment for these covariates. It showed that alteplase and tenecteplase were equivalent for 30-day mortality on both absolute and relative scales, with tenecteplase showing lower rates of some complications.

1.4.2 ASSENT-3 study

The ASSENT-3 study was an open-label randomised study that combined the new fibrinolytic therapy tenecteplase with one of three different antithrombotic conjunctive therapies: enoxaparin, abciximab or unfractionated heparin (ASSENT-3 Investigators, 2001). A total of 6095 patients with acute myocardial infarction were randomised to the study, with the main comparison of interest being between the enoxaparin ($n = 2040$) and unfractionated heparin ($n = 2038$) groups.

The study had two binary primary endpoints: efficacy (a composite of 30-day mortality, in-hospital reinfarction or in-hospital refractory ischaemia) and efficacy plus safety (the efficacy endpoint along with in-hospital major bleeding). The power calculation of the study was based on a hypothesis of non-inferiority with a 1% risk difference margin, and enoxaparin showed a clear advantage over unfractionated heparin on both primary endpoints, with unadjusted relative risks of 0.74 (95% CI 0.63–0.87) and 0.81 (0.70–0.93), respectively.

1.4.3 BOOST-NZ study

The Benefits Of Oxygen Saturation Targeting New Zealand (BOOST-NZ) study was one of a group of similar randomised trials examining the effects of in-hospital oxygen saturation targeting for premature infants (Darlow et al., 2014). Treatment was blinded by the use of specially designed oximeters which displayed a value either 3% higher or lower than the true oxygen saturation (SpO_2) at any time. Study nurses were instructed to target a displayed SpO_2 of 88–92%, such that one group received a range of 85–89% and the other received 91–95%. The true SpO_2 level at times when the infants were receiving supplementary oxygen was stored by the oximeter.

Current evidence suggests that low levels of oxygen saturation can increase the risk of death, but excessively high levels can lead to other complications, such as blindness caused by retinopathy of prematurity. The range of SpO_2 examined in the study was considered safe, but the level for optimising the trade-off between these negative outcomes was unknown. As such, the primary endpoint of the study was a composite of death or severe disability at 2 years corrected gestational age, with a secondary endpoint being 2-year mortality alone. Both of these outcomes are binary, and the

primary analyses showed relative risks of 1.15 (95% CI 0.90–1.47) and 1.10 (0.68–1.78) in favour of the low-target group for these endpoints, respectively.

1.4.4 FIELD study

The Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study was a double-blind randomised trial in 9795 patients with type 2 diabetes mellitus (FIELD Study Investigators, 2005). The aim of the study was to compare the effect of fenofibrate, a cholesterol-lowering medication, with placebo on a primary outcome of coronary events (coronary heart disease death or non-fatal myocardial infarction). The primary analysis showed an unadjusted hazard ratio of 0.89 in favour of fenofibrate, but it was not statistically significant at the pre-specified 5% level ($p = 0.16$).

With fibrates having shown potentially beneficial effects on microvascular endpoints, a secondary outcome of the FIELD study was the impact of fenofibrate on diabetic retinopathy (Keech et al., 2007). Specifically, the number of courses of laser photocoagulation therapy undertaken by each patient was a count data outcome, because patients could receive laser treatment multiple times. Fenofibrate showed a highly significant rate ratio of 0.63 (95% CI 0.49–0.81; $p = 0.0003$) compared to placebo in unadjusted analysis of this count outcome.

1.5 Thesis outline

The remainder of this thesis is organised as follows. Chapter 2 gives an overview of existing methods for semi-parametric regression modelling of rate differences, risk differences and relative risks, and their deficiencies, as well as some brief theoretical background of the methods that will be used to address these in the subsequent chapters. Chapter 3 contains a peer-reviewed journal article that describes a stable algorithm for fitting an additive binomial model in order to estimate adjusted risk differences. This is demonstrated with an application to data from the ASSENT-3 study. The method includes the possibility of semi-parametric adjustment through an isotonic step function, which we apply to data from the ASSENT-2 study. Chapter 4 contains a peer-reviewed journal article that provides an extension of the additive binomial (risk

difference) method, as well as the existing additive Poisson (rate difference) and multiplicative binomial (relative risk) methods to allow for more flexible semi-parametric adjustment. The ability to use smooth semi-parametric regression functions — with or without a monotonicity restriction — is implemented, without compromising the stability of the underlying algorithms. This is illustrated using applications of the methods to the ASSENT-2 study. In Chapter 5, a manuscript that has been prepared for submission addresses the problem of rate difference regression modelling in the presence of overdispersion, by describing a stable algorithm for fitting an additive negative binomial model. This method is applied to laser therapy data from the FIELD study. Finally, Chapter 6 provides a summary and discussion of the thesis, including a review of open-source software that implement the methodology. Future research directions based on this work are also presented.

Appendices A and B contain the documentation for two R software packages — `logbin` and `addreg` — that implement the methods described in this thesis. In Appendix C, we include a published conference proceedings paper that is an earlier version of the semi-parametric method described in Chapter 4, including an application to a different clinical trial dataset, obtained from the BOOST-NZ study.

2

Background

Methods for semi-parametric regression modelling of binary and count data are well established, and have been implemented in many widely used statistical software packages. However, the usual methods for fitting such models have some weaknesses, particularly with the non-standard models that are needed to estimate adjusted risk differences, rate differences and relative risks. In the first part of this chapter, the existing methods and their potential problems are described, and in the second part, we introduce some previous results upon which our approach will be based. Finally, we give a brief outline of the methods that will be described in detail in the following chapters.

2.1 Usual approaches and limitations

2.1.1 Generalised linear models

A generalised linear model (GLM) relates an outcome variable Y to a vector of predictors or covariates $\boldsymbol{x}$, assuming Y has one of the wide range of probability distributions

included in the exponential family (McCullagh and Nelder, 1989). In a GLM, the parameter $\lambda = \mathbb{E}(Y)$ is related to a linear combination of the covariates $\boldsymbol{x}$ through a *link function* g,

$$g(\lambda) = \boldsymbol{x}\boldsymbol{\theta}.$$

In binomial models and count data models, the distributional parameter of interest λ is the risk or rate, and the above specification can be modified to include a multiplicative standardising term N, representing the number of trials or exposure time over which Y was observed. That is, $\mathbb{E}(Y) = N\lambda$.

The linear predictor $\boldsymbol{x}\boldsymbol{\theta}$ contains an intercept, and for the methods we describe in this thesis, it is convenient to separate the covariates into a collection of A categorical and B continuous covariates: $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{v}) = (u_1, \ldots, u_A, v_1, \ldots, v_B)$, so

$$g(\lambda) = \Lambda(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_a) + \sum_{b=1}^{B} \beta_b v_b, \tag{2.1}$$

where, without loss of generality, each $u_a \in \{1, \ldots, k_a\}$ and $v_b \in \mathbb{R}$.

The parameter vector $\boldsymbol{\theta}$ contains an intercept term α_0, $\sum_a k_a$ parameters associated with the categorical covariates and B parameters associated with the continuous covariates. In order for the model parameters to be identifiable, for each categorical covariate $a = 1, \ldots, A$ we choose a reference level $r_a \in \{1, \ldots, k_a\}$ and set $\alpha_a(r_a) = 0$ so that a total of $J = 1 + \sum_{a=1}^{A}(k_a - 1) + B$ parameters must be estimated.

Keeping all other covariates constant, the parameter $\alpha_a(k)$ represents the change in $g(\lambda)$ associated with having $u_a = k$ compared to $u_a = r_a$. Likewise, β_b represents the change in $g(\lambda)$ for a one-unit increase in v_b. A simple transformation of the parameter can thus provide adjusted effect measures for each of the covariates in the model.

The choice of link function is crucial to the interpretation of these parameters, in particular with regard to the effect measures discussed in Section 1.2. The canonical link is the function that relates the canonical parameter of an exponential family distribution to its expected value. GLMs that use the canonical link generally have good mathematical properties, and so it is typically used as the default link function.

For binomial random variables where $Y \sim \text{Bin}(N, \lambda)$, the canonical link is the logit function $g(\lambda) = \log(\lambda/(1-\lambda))$ and so $\exp(\alpha_a(k))$ and $\exp(\beta_b)$ are the adjusted odds ratios associated with the categorical and continuous covariates, respectively. In order

to estimate relative risks, the log link $g(\lambda) = \log(\lambda)$ must be used, and these effect measures can be obtained by exponentiating the model parameters. Adjusted risk differences can be obtained by using the identity link $g(\lambda) = \lambda$, with no transformation required.

In a Poisson GLM, $Y \sim \text{Poisson}(N\lambda)$, where N is the exposure time and λ is the underlying event rate per unit time. The log link $g(\lambda) = \log(\lambda)$ is the canonical link, giving adjusted rate ratios when the parameters are exponentiated. Rate differences result from the identity link $g(\lambda) = \lambda$, again without needing to transform the parameters.

As explained in Section 2.1.4, a complication of these non-canonical models is that they impose implicit constraints on the parameter vector $\boldsymbol{\theta}$.

2.1.2 Maximum likelihood estimation

In order to find estimates of the parameters of interest, we use maximum likelihood estimation. Given a vector $\boldsymbol{Y}$ of n independent random outcome variables $(Y_1, \ldots, Y_n)$, where the distribution of Y_i depends on λ_i, which is related to $\boldsymbol{\theta}$ through (2.1), we wish to find

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\, \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Y}) = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\, \ell(\boldsymbol{\theta}; \boldsymbol{Y}),$$

where $\mathcal{L}$ is the likelihood function, ℓ is the log-likelihood and Θ is the parameter space which contains the valid possible values of $\boldsymbol{\theta}$. The form of $\mathcal{L}$ and ℓ will be determined by the particular exponential family distribution that is used to model $\boldsymbol{Y}$, such as the binomial or Poisson distribution.

Under certain conditions — all of which hold in the models that we consider in this thesis — the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ is asymptotically consistent for the true value $\boldsymbol{\theta}_0$ (Fahrmeir and Kaufmann, 1985). Furthermore, if $\boldsymbol{\theta}_0$ is not on the boundary of the parameter space, the MLE is asymptotically normal with

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}^{-1})$$

as $n \to \infty$, where $\boldsymbol{I}$ is the expected information matrix with (j, k) entry:

$$I_{(j,k)} = -\mathbb{E}\left\{ \frac{\partial^2}{\partial\theta_j \partial\theta_k} \ell(\boldsymbol{\theta}; \boldsymbol{Y}) \,\middle|\, \boldsymbol{\theta}_0 \right\}.$$

Thus the expected information, evaluated at the MLE, can be used to estimate standard errors and hence confidence intervals for the parameters when the MLE is not on the boundary of the parameter space. An alternative is to use the observed information matrix $\boldsymbol{I}^{(O)}$, which has (j,k) element

$$I^{(O)}_{(j,k)} = -\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ell(\boldsymbol{\theta};\boldsymbol{Y}).$$

If the MLE is on the boundary, information matrix estimates of sampling variation are invalid. In this case, confidence intervals may be estimated by using a resampling procedure such as the bootstrap. Marschner (2015) has demonstrated that such an approach gives accurate coverage probabilities for the parameters in a log-link binomial model. However, bootstrapped confidence intervals require that we have a reliable method of maximum likelihood estimation that is able to find the MLE in every bootstrap resample. As discussed in the next two sections, this is not always true of the standard methods that are commonly implemented in statistical software packages.

2.1.3 Iterative methods

It is usually not possible to derive an explicit formula for the MLE of a GLM, and instead we must use an iterative approach to maximise the likelihood. In most popular statistical software packages, either the Newton–Raphson or the Fisher scoring algorithm is implemented via an iteratively reweighted least squares (IRLS) routine. At the $(c+1)^{\text{th}}$ step, both algorithms update the current parameter estimates $\hat{\boldsymbol{\theta}}_{(c)}$ by an equation of the form

$$\hat{\boldsymbol{\theta}}_{(c+1)} = \hat{\boldsymbol{\theta}}_{(c)} - \boldsymbol{H}^{-1}\boldsymbol{s}, \tag{2.2}$$

where $\boldsymbol{s} = (s_1, \ldots s_J)^\top$ is the vector of first derivatives of the log-likelihood, that is, $s_j = \partial\ell/\partial\theta_j$, evaluated at the current estimate $\hat{\boldsymbol{\theta}}_{(c)}$. In (2.2), $\boldsymbol{H}$ is either the observed information matrix $\boldsymbol{I}^{(O)}$ (for Newtown–Raphson) or the expected information matrix $\boldsymbol{I}$ (for Fisher scoring), evaluated at $\hat{\boldsymbol{\theta}}_{(c)}$. When the canonical link function is used, the two methods are identical.

The IRLS routine derives from the fact that for GLMs, $\boldsymbol{H}$ can be written in the form

$$\boldsymbol{H} = -\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X},$$

where $\boldsymbol{X}$ is the $n \times J$ design matrix associated with the model and $\boldsymbol{W}$ is a diagonal matrix with elements that depend on the outcome distribution and link function. Similarly, $\boldsymbol{s}$ can be expressed as $\boldsymbol{X}^\top \boldsymbol{W} \widetilde{\boldsymbol{Y}}$ with the same $\boldsymbol{W}$ and a transformed outcome vector $\widetilde{\boldsymbol{Y}}$. Then the parameter estimate update (2.2) is

$$\hat{\boldsymbol{\theta}}_{(c+1)} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Z},$$

where $Z_i = \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}_{(c)}) + \widetilde{Y}_i$. This is the solution of a weighted least squares regression problem, where $\boldsymbol{Z}$ and $\boldsymbol{W}$ must be recomputed at each step and the process is repeated until convergence of the parameter estimates.

A typical way of defining convergence is to use

$$\frac{\|\hat{\boldsymbol{\theta}}_{(c+1)} - \hat{\boldsymbol{\theta}}_{(c)}\|}{\|\hat{\boldsymbol{\theta}}_{(c)}\|} < \epsilon \tag{2.3}$$

for some small positive constant ϵ. This convergence criterion will be used for all of the algorithms defined in this thesis.

2.1.4 Convergence issues

The iterative methods described in Section 2.1.3 work well for canonical models but they can be subject to convergence problems for the non-canonical models described in Section 2.1.1. Unlike the canonical models, these non-canonical models impose implicit parameter constraints that complicate the convergence behaviour. This arises because the iterative methods are based on the gradient of the score function, which can cause the updated estimates to overshoot the maximum and leave the constrained parameter space. In this section, we discuss these issues.

In binomial models, the range of the log-odds is unrestricted, as any value will correspond to a valid risk $\lambda \in [0, 1]$. Similarly for count models, the rate λ must be non-negative, so $\log(\lambda)$ can take any value. Thus when the canonical link is used with these models, there is no constraint on the parameter space; that is, $\Theta = \mathbb{R}^J$.

With the other link functions described in Section 2.1.1, however, Θ must be constrained such that the fitted model will produce only valid risks and rates. Specifically,

for the log-link binomial model, the parameter space

$$\Theta = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) \leq 0, (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}\} \tag{2.4}$$

ensures that fitted risks are in $[0, 1]$, where $\mathcal{U} \times \mathcal{V}$ is the Cartesian product of the covariate spaces corresponding to $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. We will take these covariate spaces to be the Cartesian product of the observed ranges of the covariates,

$$\mathcal{U} = \prod_{a=1}^{A} \{1, \ldots, k_a\} \qquad \text{and} \qquad \mathcal{V} = \prod_{b=1}^{B} [v_b^{(0)}, v_b^{(1)}], \tag{2.5}$$

where $v_b^{(0)} = \min_i\{v_{ib}\}$ and $v_b^{(1)} = \max_i\{v_{ib}\}$.

Likewise, for the identity-link binomial model, we have

$$\Theta = \{\boldsymbol{\theta} : 0 \leq \Lambda(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) \leq 1, (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}\}, \tag{2.6}$$

and for the identity-link count data model

$$\Theta = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) \geq 0, (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}\}. \tag{2.7}$$

The Newton–Raphson and Fisher scoring algorithms have no inbuilt mechanism for these constrained optimisation problems, and it is possible for the estimates to move outside the parameter space at any iteration. A common method for attempting to resolve this is to use step-halving: if $\hat{\boldsymbol{\theta}}^*_{(c+1)}$ is the (invalid) proposed update to the vector of parameter estimates, it is replaced by

$$\hat{\boldsymbol{\theta}}_{(c+1)} = \hat{\boldsymbol{\theta}}_{(c)} + \frac{1}{2^d}\left(\hat{\boldsymbol{\theta}}^*_{(c+1)} - \hat{\boldsymbol{\theta}}_{(c)}\right),$$

for $d = 1, 2, \ldots$, until the new $\hat{\boldsymbol{\theta}}_{(c+1)}$ is inside the parameter space. This too is not guaranteed to yield the MLE, particularly if the MLE is on the boundary of the parameter space and the gradient of the log-likelihood is almost perpendicular to the boundary, in which case convergence may be declared before the MLE is reached (Lumley, Kronmal, and Ma, 2006). Problems can also occur even when the MLE is a stationary point in the interior of the parameter space. Examples of such behaviour are provided in

subsequent chapters.

2.1.5 Generalised additive models

In some situations, we may wish to relax the modelling assumption that the relationship between $g(\lambda)$ and each of the continuous covariates is strictly linear. Generalised additive models (GAMs) were introduced by Hastie and Tibshirani (1990), and provide this additional flexibility by introducing unspecified functions f_c, $c = 1, \ldots, C$ into the linear predictor (2.1), so

$$g(\lambda) = \Lambda(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}; \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_a) + \sum_{b=1}^{B} \beta_b v_b + \sum_{c=1}^{C} f_c(w_c).$$

There are a number of ways in which this can be achieved, and in this thesis we will employ regression splines. The unknown f_c are assumed to belong to a class of functions that can be expressed in terms of a set of D_c basis functions B_{cd}, such that

$$f_c(w) = \sum_{d=1}^{D_c} \gamma_{cd} B_{cd}(w), \tag{2.8}$$

and the problem becomes one of estimating the γ_{cd} parameters. Details of the spline basis functions used in this thesis are provided in Sections 2.2.4 and 2.2.5 below.

Software packages implement GAMs in different ways: a summary of these is given in Section 4.6. All of them, however, use an iterative algorithm that is some variant of the Newton–Raphson or Fisher scoring algorithm and are hence subject to the same instability that can occur with GLMs, particularly for the constrained models that we consider in this thesis.

2.2 Theoretical background

The following sections provide more theoretical details of the important results that are used but only introduced briefly in the later chapters of this thesis.

2.2.1 Multinomial–Poisson transformation

The multinomial–Poisson transformation (Baker, 1994) is a method for performing maximum likelihood estimation for a multinomial model by maximising a likelihood function associated with a Poisson model.

Let $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ij}, \ldots)$, $i = 1, \ldots, n$, be observations from a multinomial distribution with parameters $g_{ij}(\boldsymbol{\theta})/G_i(\boldsymbol{\theta})$ for $j \in J_i$, where J_i is the set of possible outcomes for individual i and

$$G_i(\boldsymbol{\theta}) = \sum_{j \in J_i} g_{ij}(\boldsymbol{\theta}).$$

After removing terms that do not depend on the parameter vector $\boldsymbol{\theta}$, the likelihood for the multinomial model is

$$\mathcal{L}_{\mathrm{M}}(\boldsymbol{\theta}; \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{j \in J_i} \left(\frac{g_{ij}(\boldsymbol{\theta})}{G_i(\boldsymbol{\theta})} \right)^{Z_{ij}}. \tag{2.9}$$

We introduce n new parameters $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)$, restricting each $\varphi_i > 0$. The multinomial–Poisson transformation of (2.9) is

$$\mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta}, \boldsymbol{\varphi}; \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{j \in J_i} (\varphi_i g_{ij}(\theta))^{Z_{ij}} \exp(-\varphi g_{ij}(\theta)). \tag{2.10}$$

For fixed $\boldsymbol{\theta}$, it is easy to find the maximum likelihood estimate of $\boldsymbol{\varphi}$, denoted by $\hat{\boldsymbol{\varphi}}(\boldsymbol{\theta}) = (\hat{\varphi}_1(\boldsymbol{\theta}), \ldots, \hat{\varphi}_n(\boldsymbol{\theta}))$. With $N_i = \sum_j Z_{ij}$,

$$\hat{\varphi}_i(\boldsymbol{\theta}) = \frac{N_i}{G_i(\boldsymbol{\theta})}.$$

Substituting this back into (2.10) gives

$$\begin{aligned} \mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}(\boldsymbol{\theta}); \boldsymbol{Z}) &= \left(\prod_{i=1}^{N} N_i^{N_i} \exp(N_i) \right) \mathcal{L}_{\mathrm{M}}(\boldsymbol{\theta}; \boldsymbol{Z}) \\ &\propto \mathcal{L}_{\mathrm{M}}(\boldsymbol{\theta}; \boldsymbol{Z}). \end{aligned}$$

It follows from the work of Richards (1961) on profile likelihoods that the value of $\boldsymbol{\theta}$ that maximises $\mathcal{L}_{\mathrm{P}}$ is the same as that which maximises $\mathcal{L}_{\mathrm{M}}$. This means that in order to find the MLE from a multinomial model, we can instead find the MLE associated

with $\mathcal{L}_{\mathrm{P}}$ which, up to a multiplicative constant, takes the same form as the likelihood for a Poisson model where

$$Z_{ij} \sim \text{Poisson}(\varphi_i g_{ij}(\boldsymbol{\theta})).$$

We will make use of this for the binomial special case of the multinomial distribution, where the equivalent Poisson likelihood is easier to work with than the actual binomial likelihood.

2.2.2 Expectation–Maximisation (EM) algorithm and variants

The Expectation–Maximisation (EM) algorithm (Dempster, Laird, and Rubin, 1977) is a method for performing maximum likelihood estimation with incomplete data. Beginning with an initial set of parameter estimates, the algorithm uses these to 'fill in' the missing observations, and updates the estimates by maximising the likelihood based on the hypothetical complete data. Each of these iterations increases the likelihood of the observed data monotonically, and the process is repeated until convergence is achieved.

A large number of variations of the basic EM algorithm have been described, providing faster convergence or simpler mathematical derivations in some situations (McLachlan and Krishnan, 1997). We give details of some of these below.

EM algorithm

The concept of 'incomplete data' in the context of an EM algorithm refers to the situation in which we have an observed data vector $\boldsymbol{Y}$ that is associated with an unobserved data vector $\boldsymbol{\mathcal{Y}}$ through a many-to-one mapping from the sample space of $\boldsymbol{\mathcal{Y}}$ to the sample space of $\boldsymbol{Y}$. The 'complete data' $\boldsymbol{\mathcal{Y}}$ is not observed directly, but only indirectly through $\boldsymbol{Y}$.

We wish to find the MLE of the parameter vector $\boldsymbol{\theta} \in \Theta$ based on the log-likelihood $\ell(\boldsymbol{\theta}; \boldsymbol{Y})$ for the observed data. The EM algorithm is useful in situations where this maximum likelihood estimation would be straightforward if the complete data $\boldsymbol{\mathcal{Y}}$ was available, in which case we could instead maximise the complete-data log-likelihood $L(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}})$.

The basic EM algorithm is made up of alternating E- and M-steps, iterated until

convergence. Given an initial value for the parameter estimate $\hat{\boldsymbol{\theta}}_{(0)}$, the E-step at the $(c+1)^{\text{th}}$ iteration requires calculation of

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{(c)}) = \mathbb{E}\left(L(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}) \mid \boldsymbol{Y}, \hat{\boldsymbol{\theta}}_{(c)}\right). \tag{2.11}$$

The M-step then involves maximisation of Q with respect to $\boldsymbol{\theta}$, so the updated parameter estimate is

$$\hat{\boldsymbol{\theta}}_{(c+1)} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\, Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{(c)}).$$

The EM algorithm ensures that the likelihood will never decrease between iterations, and because the M-step only considers estimates in the parameter space, this approach can never lead to iterates outside the parameter space, as can occur with the gradient-based algorithms described in Section 2.1.3. This makes the EM algorithm a potentially useful approach in constrained estimation problems such as those considered in this thesis.

Wu (1983) has shown that if the log-likelihood is unimodal in the interior of the parameter space Θ and $\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')/\partial \boldsymbol{\theta}$ is continuous in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, then the sequence of estimates $\{\hat{\boldsymbol{\theta}}_{(c)}\}$ from an EM algorithm will converge to the unique MLE. Unimodality of ℓ holds for exponential family distributions because the log-likelihood function is globally concave with respect to $\boldsymbol{\theta}$.

ECM algorithm

The Expectation–Conditional Maximisation (ECM) algorithm (Meng and Rubin, 1993) is an extension of the EM algorithm in which the M-step of each iteration is replaced by a sequence of D conditional maximisation (CM) steps. Each CM-step maximises Q with respect to $\boldsymbol{\theta}$, but keeps a subset of the parameters $h_d(\boldsymbol{\theta})$ $(d = 1, \ldots, D)$ fixed at their current estimates.

At the $(c+1)^{\text{th}}$ iteration, the E-step remains as defined by (2.11). The first CM-step updates the current estimate by maximising $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{(c)})$ subject to the constraint $h_1(\boldsymbol{\theta}) = h_1(\hat{\boldsymbol{\theta}}_{(c)})$, giving $\hat{\boldsymbol{\theta}}_{(c+1/D)}$. The second CM-step requires maximisation of $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{(c)})$ subject to $h_2(\boldsymbol{\theta}) = h_2(\hat{\boldsymbol{\theta}}_{(c+1/D)})$, to get $\hat{\boldsymbol{\theta}}_{(c+2/D)}$. This is repeated for $d = 3, \ldots, D$, where, in general

$$\hat{\boldsymbol{\theta}}_{(c+d/D)} = \underset{\boldsymbol{\theta} \in \Theta(c,d)}{\arg\max}\, Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{(c)}), \tag{2.12}$$

with $\Theta(c,d) = \{\boldsymbol{\theta} \in \Theta : h_d(\boldsymbol{\theta}) = h_d(\hat{\boldsymbol{\theta}}_{(c+(d-1)/D)})\}$, and the whole sequence of E- and CM-steps is repeated until convergence.

The ECM algorithm is useful in situations where it is easier to maximise Q with respect to particular subsets $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$ of the parameter vector separately, instead of considering the entire $\boldsymbol{\theta}$ at once, as is required by the M-step of the EM algorithm. In such a case, each $h_d(\boldsymbol{\theta})$ would consist of all of the elements in $\boldsymbol{\theta}$ except for $\boldsymbol{\theta}_d$.

Meng and Rubin (1993) showed that the ECM algorithm has similar convergence properties to the EM algorithm, with monotone convergence to the MLE in the interior of the parameter space, subject to the additional condition that the set $H = \{h_d(\boldsymbol{\theta}); d = 1, \ldots, D\}$ is 'space-filling'. Intuitively, this condition means that from any point in Θ, one is able to search in any direction for the maximum, so that the resulting maximisation is over the original Θ rather than a subspace of it.

ECME algorithm

The ECME algorithm (Liu and Rubin, 1994) is a generalisation of the ECM algorithm. Standing for Expectation–Conditional Maximisation–Either, the ECME algorithm allows some of the CM-steps of the ECM algorithm, which maximise the expected complete-data log-likelihood, to be replaced by steps that maximise the observed-data log-likelihood.

That is, the E-step is the same as for the EM and ECM algorithms, and we define D parameter subsets $\{h_1(\boldsymbol{\theta}), \ldots, h_D(\boldsymbol{\theta})\}$, as for the ECM algorithm. Then $\mathcal{D} = \{1, \ldots, D\}$ is partitioned into two sets, $\mathcal{D}_Q$ and $\mathcal{D}_L$, such that within the $(c+1)^{\text{th}}$ iteration, if $d \in \mathcal{D}_Q$, we update the parameter estimates using (2.12), and for $d \in \mathcal{D}_L$, we use

$$\hat{\boldsymbol{\theta}}_{(c+d/D)} = \underset{\boldsymbol{\theta} \in \Theta(c,d)}{\arg\max}\, \ell(\boldsymbol{\theta}; \boldsymbol{Y}).$$

If the $\mathcal{D}_L$ steps in an iteration are performed after all of the $\mathcal{D}_Q$ steps have been completed, this algorithm shares similar convergence properties to the EM and ECM algorithms, requiring only the additional condition that each of the conditional maximisations within an iteration are unique (Meng and Van Dyk, 1997).

2.2.3 Combinatorial EM algorithms

The EM algorithm and its variants lend themselves to situations in which the outcome variable can be represented as a function of a collection of unobserved latent outcome variables. These underlying outcome variables can be thought of as the missing data in an EM-type formulation. This is the manner in which we will use the EM algorithm and its variants in this thesis.

We will see that when the usual gradient-based algorithms fail to converge to the MLE, an EM-type algorithm can provide stable monotone convergence, while still respecting the required parameter space constraints.

The parameter spaces for the models considered in this thesis, defined in (2.4), (2.6) and (2.7), do not impose constraints on the individual parameters, only on the resulting linear predictor. However, due to the way in which we will define the latent outcome variables, an EM algorithm will apply additional constraints to the individual model parameters. Modifications are therefore needed so that the entire parameter space Θ is searched for the MLE. This is the basic idea behind the combinatorial EM (CEM) algorithm described by Marschner (2014). We give two examples of applications of CEM algorithms below, before giving a more general description. In subsequent chapters, our methods will be based on implementations of CEM algorithms.

Additive Poisson model

We will refer to the identity-link Poisson GLM as an additive Poisson model. In an additive Poisson model, we assume each $Y_i \sim \text{Poisson}(N_i\lambda_i)$, $i = 1, \ldots, n$, where the standardised event rate is $\lambda_i = \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \boldsymbol{\theta})$, as defined in (2.1). The parameters in this model represent adjusted rate differences, and the parameter space (2.7) ensures that the fitted rates are non-negative.

A CEM algorithm for additive Poisson regression has been described by Marschner (2010). We begin by choosing $\boldsymbol{r} = (r_1, \ldots, r_A)$, where each $r_a \in \{1, \ldots, k_a\}$, imposing the necessary identifiability constraints on the categorical parameters. For the linear continuous covariates, we also choose $\boldsymbol{s} = (s_1, \ldots, s_B)$, where each $s_b = v_b^{(\varsigma_b)}$ for a choice of $\varsigma_b \in \{0, 1\}$. Given $\boldsymbol{s}$, we define the transformed covariates $v'_{ib} = (-1)^{\varsigma_b}(v_{ib} - s_b)$ for $b = 1, \ldots, B$, which are always non-negative. Then the rate can be equivalently

expressed as

$$\begin{aligned}\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \boldsymbol{\theta}) &= \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i'; \boldsymbol{\theta}') \\ &= \alpha_0' + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta_b' v_{ib}',\end{aligned}$$

where

$$\alpha_0' = \alpha_0 + \sum_{b=1}^{B} \beta_b s_b \qquad \text{and} \qquad \beta_b' = (-1)^{\varsigma_b}\beta_b. \tag{2.13}$$

Because the sum of independent Poisson random variables also has a Poisson distribution, Y_i can be thought of as the sum of $1 + A + B$ latent outcome variables, that is

$$Y_i = \mathcal{Y}_i^{(0)} + \sum_{a=1}^{A} \mathcal{Y}_i^{(a)} + \sum_{b=1}^{B} \mathcal{Y}_i^{(A+b)},$$

where

$$\begin{aligned}\mathcal{Y}_i^{(0)} &\sim \text{Poisson}(N_i\alpha_0') \\ \mathcal{Y}_i^{(a)} &\sim \text{Poisson}(N_i\alpha_a(u_{ia})) \qquad a = 1, \ldots, A \\ \mathcal{Y}_i^{(A+b)} &\sim \text{Poisson}(N_i\beta_b' v_{ib}') \qquad b = 1, \ldots, B.\end{aligned}$$

To find the MLE, an EM algorithm can be defined in which the unobserved $\mathcal{Y}_i^{(j)}$, $i = 1, \ldots, n$, $j = 0, \ldots, A + B$, are treated as missing data. However, because the means of these latent Poisson variables must be non-negative, this complete-data model imposes a non-negativity constraint on each of the individual model parameters. Given non-negative starting estimates $\hat{\boldsymbol{\theta}}'^{(0)} = (\hat{\alpha}_0'^{(0)}, \hat{\boldsymbol{\alpha}}_1^{(0)}, \ldots, \hat{\boldsymbol{\alpha}}_A^{(0)}, \hat{\beta}_1'^{(0)}, \ldots \hat{\beta}_B'^{(0)})$, the updated estimates at the $(c+1)^{\text{th}}$ iteration are given by

$$\begin{aligned}\hat{\alpha}_0'^{(c+1)} &= \hat{\alpha}_0'^{(c)} \frac{\sum_{i=1}^{n} \left(Y_i / \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}) \right)}{\sum_{i=1}^{n} N_i} \\ \hat{\alpha}_a^{(c+1)}(u) &= \hat{\alpha}_0^{(c)}(u) \frac{\sum_{i \in I_{au}} \left(Y_i / \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}) \right)}{\sum_{i \in I_{au}} N_i} \\ \hat{\beta}_b'^{(c+1)} &= \hat{\beta}_b'^{(c)} \frac{\sum_{i=1}^{n} \left(Y_i v_{ib}' / \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}) \right)}{\sum_{i=1}^{n} N_i v_{ib}'},\end{aligned}$$

where $I_{au} = \{i : u_{ia} = u\}$, and the estimates are guaranteed to remain non-negative at

every iteration.

In fact, this non-negativity restriction on each of the parameters ensures that the covariate vector $(\boldsymbol{r}, \boldsymbol{s}) = (r_1, \ldots, r_A, s_1, \ldots, s_B)$ gives the smallest fitted rate of any $(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}$. That is, the parameter space considered by the EM algorithm for a particular choice of $(\boldsymbol{r}, \boldsymbol{s})$ is

$$\Theta(\boldsymbol{r}, \boldsymbol{s}) = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) \geq \Lambda(\boldsymbol{r}, \boldsymbol{s}; \boldsymbol{\theta}) \geq 0, (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}\}.$$

The key to the CEM algorithm is that the union of these parameter subspaces coincides with the full parameter space Θ. This means that if we implement the EM algorithm for each possible choice of $(\boldsymbol{r}, \boldsymbol{s})$, we will find a constrained MLE for each parameter subspace, and that which has the highest likelihood must be the overall MLE. The total number of parameter subspaces is $\prod_{a=1}^{A} k_a \times 2^B$.

Once we have found the overall MLE, we invert (2.13) to obtain the parameter estimates on their original scale.

Log-link binomial model

For the log-link binomial model, we have $Y_i \sim \text{Bin}(N_i, \lambda_i)$ where $\log(\lambda_i) = \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \boldsymbol{\theta})$. The fitted risk λ_i must be in $[0, 1]$, and so the linear predictor requires a non-positivity constraint, as defined in (2.4).

The CEM algorithm for this model, presented by Marschner and Gillett (2012), is based on the fact that the product of independent Bernoulli random variables is also Bernoulli, with an event probability that is the product of the individual event probabilities. Thus the EM algorithm is derived by considering the binary outcome to be composed of a collection of unobserved Bernoulli random variables.

The parameter space is partitioned in a similar way as for the additive Poisson model, except here the latent event probabilities must be in $[0, 1]$, so the complete-data model imposes non-positivity constraints on the individual parameters in the model. This corresponds to a parameter subspace of

$$\Theta(\boldsymbol{r}, \boldsymbol{s}) = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) \leq \Lambda(\boldsymbol{r}, \boldsymbol{s}; \boldsymbol{\theta}) \leq 0, (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}\},$$

that is, the space in which the reference vector $(\boldsymbol{r}, \boldsymbol{s})$ corresponds to the largest fittest

risk. In a similar fashion to the additive Poisson model, it is easy to verify that the union of these subspaces over all possible choices of $(\boldsymbol{r}, \boldsymbol{s})$ is the full parameter space Θ.

At the $(c+1)^{\text{th}}$ iteration of the EM algorithm, the parameter estimates are updated using

$$\begin{aligned}
\hat{\alpha}_0'^{(c+1)} &= \log\left(\sum_{i=1}^{n} \hat{\mathcal{Y}}_i^{(0)(c)} \Big/ \sum_{i=1}^{n} N_i\right) \\
\hat{\alpha}_a^{(c+1)}(u) &= \log\left(\sum_{i \in I_{au}} \hat{\mathcal{Y}}_i^{(a)(c)} \Big/ \sum_{i \in I_{au}} N_i\right) \\
\hat{\beta}_b'^{(c+1)} &= \log\left(\sum_{i=1}^{n} \hat{\mathcal{Y}}_i^{(A+b)(c)} v'_{ib} \Big/ \sum_{i=1}^{n} N_i v'_{ib}\right),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mathcal{Y}}_i^{(0)(c)} &= Y_i + (N_i - Y_i)\frac{\exp(\hat{\alpha}_0'^{(c)}) - \exp(\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}))}{1 - \exp(\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}))} \\
\hat{\mathcal{Y}}_i^{(a)(c)} &= Y_i + (N_i - Y_i)\frac{\exp(\hat{\alpha}_a^{(c)}(u_{ia})) - \exp(\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}))}{1 - \exp(\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}))} \\
\hat{\mathcal{Y}}_i^{(A+b)(c)} &= Y_i + (N_i - Y_i)\frac{\exp(\hat{\beta}_b'^{(c)}) - \exp(\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}))}{1 - \exp(\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i; \hat{\boldsymbol{\theta}}^{(c)}))}.
\end{aligned}$$

Starting with non-positive initial estimates, the updated parameter estimates are guaranteed to remain non-positive at all iterations, and hence are always in the parameter subspace $\Theta(\boldsymbol{r}, \boldsymbol{s})$. The EM algorithm is iterated until convergence, and repeated to find the constrained MLE for each parameter subspace. Again, the overall MLE is the constrained MLE with the highest likelihood and we can return the estimates to their original scale by inverting (2.13).

General CEM algorithm

The previous examples, and the new methods developed in this thesis, are implementations of CEM algorithms (Marschner, 2014). In general, a CEM algorithm is an approach in which we consider a finite family of complete-data models, indexed by $t \in \mathcal{T}$, each of which has a parameter space $\Theta(t)$ that is a subset of the parameter

space Θ for the model of interest, such that

$$\bigcup_{t \in \mathcal{T}} \Theta(t) = \Theta. \tag{2.14}$$

The complete-data models are defined such that an EM algorithm can be used to find the constrained maximum likelihood estimate $\hat{\boldsymbol{\theta}}(t)$ within each $\Theta(t)$. Then, due to (2.14), the $\hat{\boldsymbol{\theta}}(t)$ that attains the highest likelihood is the MLE $\hat{\boldsymbol{\theta}}$.

In fact, any generalised EM algorithm (Dempster, Laird, and Rubin, 1977) can be used to find each $\hat{\boldsymbol{\theta}}(t)$, and will provide stable convergence under the conditions mentioned in Section 2.2.2. In the examples given above, the standard EM algorithm was used, while Marschner, Gillett, and O'Connell (2012) employed the ECM algorithm to fit a stratified additive Poisson model. In Chapter 5, we use an ECME algorithm for fitting an additive negative binomial model.

For models with a large number of predictors, an exhaustive search over all parameter subspaces can be computationally expensive, particularly because an EM algorithm must be applied within each. However, the calculations within each subspace are independent, and so the total computing time may be reduced by running the EM algorithms in parallel on multiple CPUs.

There are also some strategies that can be employed to reduce the number of parameter subspaces that must be searched for the global maximum (Marschner, 2014). One of the simplest is that when the log-likelihood function is globally concave, stationarity is a sufficient condition for an estimate to be the overall MLE. Thus, if a stationary maximum is found within a particular subspace, this is the overall MLE and no further subspaces need to be searched.

2.2.4 B-splines

Marschner (2010) and Marschner and Gillett (2012) extended the additive Poisson and log-link binomial GLMs to GAMs by effectively using step functions as the set of basis functions in (2.8), such that the model parameters represent the increments in the unknown f_c between observed values of the covariate w_c. If these increments are allowed to take any value, the model can be fitted by simply treating the unique observed values as the levels of a categorical covariate. Alternatively, if the relationship between the

covariate and the risk or rate can be assumed to be monotonic, the increments can be restricted to be either non-negative or non-positive by exploiting the natural constraints applied by the underlying complete-data models. In Section 3.5, we use this approach to extend the additive binomial model in a similar way.

However, although this method provides some flexibility for continuous covariates beyond a linear relationship, it typically requires that a large number of parameters are estimated, and, in the case of biological processes that produce a smooth relationship between a continuous covariate and risk, the resulting jagged regression curve may not be an accurate representation of reality. In Chapter 4, we demonstrate that an alternative set of basis functions, the B-splines, can be used to provide smooth flexible regression without requiring a large number of degrees of freedom.

The B-splines (de Boor, 1978) are a family of polynomial splines constructed from the B-spline basis functions. Any polynomial spline is made up of a series of polynomials joined end-to-end at a series of q fixed turning points $\xi_1 < \cdots < \xi_q$, where ξ_1 and ξ_q are the endpoints of the range of the continuous covariate. Between any two adjacent turning points, the spline is a polynomial of order κ (degree $\kappa - 1$). Usually, it is desirable that the polynomials join smoothly at the turning points, but the B-splines can be defined to allow for different levels of smoothness at the turning points, or even discontinuities.

To achieve this, we use the notation of Ramsay (1988) and specify the desired continuity criteria by using $\nu_d \leq \kappa$ $(d = 1, \ldots, q-2)$ to represent the case where the curves meeting at ξ_{d+1} agree up to their $(\nu_d - 1)^{\text{th}}$ derivative. That is, if $\nu_d = 2$, the adjacent curves will have matching gradients at the turning point; if $\nu_d = 1$, they meet at the same value but need not have any matching derivatives; if $\nu_d = 0$, the resulting spline is allowed to be completely discontinuous at ξ_{d+1}.

The full specification of the spline can be made by defining a sequence of *knots* $\tau_1 \leq \cdots \leq \tau_{D+\kappa}$, where D is the number of free parameters that are required to specify the resulting spline: $D = (q-1)\kappa - \sum_d \nu_d$. We begin by placing κ knots at both the lower and upper ends of the range of the covariate, so $\tau_1 = \cdots = \tau_\kappa = \xi_1$ and $\tau_{D+1} = \cdots = \tau_{D+\kappa} = \xi_q$. The number of knots placed at each of the internal turning points depends on the desired level of continuity at that point. Specifically, we place $\kappa - \nu_{e-1}$ knots at each ξ_e $(e = 2, \ldots, q-1)$.

In most cases, it is desirable for the component polynomials of order κ to agree up to their $(\kappa - 1)^{\text{th}}$ derivative, which is achieved by placing a single knot at each of the internal turning points, giving $2(\kappa - 1) + q$ knots in total. The number of free parameters is then $\kappa + q - 2$, that is, the number of internal turning points plus the order of the component polynomials.

Given a set of knots $\boldsymbol{\tau}$, the D B-spline basis functions of order κ can be defined recursively as

$$B_d(w \mid \kappa) = \begin{cases} 1 & \text{if } w \in [\tau_d, \tau_{d+1}) \\ 0 & \text{otherwise} \end{cases}$$

for $\kappa = 1$, and

$$B_d(w \mid \kappa) = \frac{w - \tau_d}{\tau_{d+\kappa-1} - \tau_d} B_d(w \mid \kappa - 1) + \frac{\tau_{d+\kappa} - w}{\tau_{d+\kappa} - \tau_{d+1}} B_{d+1}(w \mid \kappa - 1)$$

for $\kappa > 1$. This is implemented in the R function `splineDesign` in the `splines` package (R Core Team, 2013).

The B-splines of order 3 with one internal turning point ($q = 3$) are shown in Figure 4.1(A). Each basis function $B_d(w \mid \kappa)$ is non-negative for $w \in [\tau_d, \tau_{d+\kappa})$ and zero elsewhere, such that if all of the B-spline coefficients are non-negative, the resulting function will be non-negative for all w, and each coefficient has only local influence on the shape of the smooth curve. This is demonstrated with an example in Section 4.4.2. Furthermore, the B-spline bases are normalised such that $\sum_d B_d(w) = 1$ for all w.

2.2.5 I-splines

Ramsay (1988) also defines a set of I-spline basis functions, which are so named because they are obtained by integrating the B-spline bases:

$$I_d(w \mid \kappa) = \frac{\kappa}{\tau_{d+\kappa} - \tau_d} \int_{\xi_1}^{w} B_d(w \mid \kappa) \, \mathrm{d}w.$$

In the case where we have the maximum level of (non-trivial) continuity, with each $v_d = \kappa - 1$ and so only one knot at each turning point, this can be written in the

equivalent form

$$I_d(w \mid \kappa) = \begin{cases} 0 & \text{if } w < \tau_d \\ \sum_{e=d}^{f} B_e(w \mid \kappa + 1) & \text{if } w \in [\tau_f, \tau_{f+1}) \quad \text{for} \quad f = d, \ldots, d + \kappa \\ 1 & \text{if } w \geq \tau_{d+\kappa} \end{cases}$$
$$= \sum_{e=d}^{D} B_e(w \mid \kappa + 1),$$

where $D = \kappa + q - 1$ and the second line follows because $B_e(w \mid \kappa + 1)$ is 0 outside $w \in [\tau_e, \tau_{e+\kappa+1})$. Note that here, the κ used in the notation of the I-splines refers to the order of the integrated B-splines, and hence the order of the I-spline polynomial is actually $\kappa+1$. Furthermore, it is easy to verify that the first I-spline, $I_1(w)$, is equal to 1 for all w due to the normalisation of the B-splines. I-splines are useful because if their coefficients are all non-negative, the resulting curve will be monotonically non-decreasing. We use such an approach in Section 4.4.5 to impose a monotonicity constraint on the smooth semi-parametric regression function, which is motivated by the work of Leitenstorfer and Tutz (2007) and Tutz and Leitenstorfer (2007).

2.2.6 Model selection

The choice of basis functions places some restrictions on the shape of the estimated semi-parametric function; for example, the location of possible turning points. Maximum flexibility can be achieved by allowing for a large number of turning points, but this can lead to overfitting. To remedy this, Eilers and Marx (1996) introduced penalised B-splines (P-splines), where a penalty term is subtracted from the log-likelihood such that large fluctuations in the fitted curve are discouraged (Green and Silverman, 1994).

However, any sensible penalty term will lose the parameter separation that is crucial to the stability of the EM algorithm, turning the M-step into a multidimensional root-finding problem (Marschner and Gillett, 2012, Supplementary materials). Instead we will use unpenalised maximum likelihood estimation and focus on choosing the optimal number of turning points by using a model selection criterion.

The *AIC* (Akaike, 1974) is an estimate of the Kullback–Leibler divergence between

two models. In particular, it is used to estimate the information lost when a model f is used to approximate the unknown 'full reality' f^*. It is defined as

$$AIC = 2J - 2\ell,$$

where ℓ is the maximised log-likelihood for the model based on the observed data, and J is a bias-adjustment term, acting as an estimate of model complexity (Burnham and Anderson, 2002, pp. 60–64). In some generalised additive models, the computation of J requires complex calculations, but for unpenalised maximum likelihood estimation, it is simply the number of estimable parameters in the approximating model f (Wood, 2006, pp. 170–171).

The AIC_c (Sugiura, 1978) includes a second-order bias correction term for small sample sizes, and is defined as

$$\begin{aligned} AIC_c &= \frac{2Jn}{n - J - 1} - 2\ell \\ &= AIC + \frac{2J(J+1)}{n - J - 1}. \end{aligned}$$

For large n, it is practically identical to AIC. Although its derivation is based on a model with a normally distributed outcome, it has shown good empirical performance for non-normal outcomes (Simonoff, 2003), and Burnham and Anderson (2002, pp. 66, 328) recommend that it is used instead of AIC for comparing models in which J is large relative to n, as long as sample observations are independent and the underlying distribution is unimodal and not heavily skewed.

Out of a set of candidate models with different numbers of turning points, the one with the smallest AIC or AIC_c may be selected as the best. The AIC and AIC_c do not require that the candidate models are nested, so the placement of the turning points is flexible in this regard. These criteria can also be used to compare models that propose different distributional forms for the outcome variable.

2.3 Method outline

We end this chapter with an outline of the three main contributions of this thesis, which are presented in detail in subsequent chapters.

2.3.1 Additive binomial model

In order to estimate adjusted risk differences, we use a binomial GLM with an identity link function, which we refer to as an additive binomial model. Our method for fitting additive binomial models makes use of the multinomial–Poisson transformation, outlined in Section 2.2.1. The binomial distribution is a special case of the multinomial, and given an additive model for the risk, we can define an equivalent additive Poisson model that will have the same MLE. After making the appropriate transformations to the data, we can therefore adapt the method of Marschner (2010) outlined in Section 2.2.3 in order to reliably find the MLE of the additive Poisson model, and then back-transform it to obtain the additive binomial MLE. Our method also allows for semi-parametric regression by including flexible isotonic relationships between covariates and the risk.

2.3.2 Flexible semi-parametric regression

We present a general method for incorporating smooth semi-parametric regression into the additive Poisson, log-link binomial and additive binomial models by using B-splines, introduced in Section 2.2.4. The properties of the B-spline basis functions, along with the fact that it is straightforward to impose non-negativity or non-positivity constraints on chosen parameters in the CEM algorithms used to fit these models, mean that they can be included by extending the latent outcome model underlying the CEM algorithm and searching the appropriate parameter subspaces for the MLE. It is also straightforward to impose a monotonicity restriction on the smooth curves, using the I-splines described in Section 2.2.5.

2.3.3 Additive negative binomial model

The method for fitting an additive negative binomial regression model follows a similar idea to that used for the CEM algorithms described in Section 2.2.3. We consider

an additive negative binomial model in which the outcome variable can be viewed as the sum of independent latent variables, and this complete-data model imposes some additional parameter constraints. For a particular parameter subspace, we define an ECME algorithm to find the constrained MLE of the parameters that represent adjusted rate differences, as well as the coefficient of overdispersion, and search over the full series of parameter subspaces in order to find the global MLE. This model can also include semi-parametric regression functions, with or without monotonicity restrictions.

3

Additive binomial regression

As discussed in Chapter 1, risk is the fundamental biostatistical measure associated with binary outcomes, and the absolute risk difference is a useful measure of the effect of an intervention or risk factor. Adjusted risk differences can be estimated using a binomial generalised linear model with an identity link function, which we refer to as an additive binomial model. However, as described in Chapter 2, the natural constraints placed on the parameter space can mean that it is difficult for standard algorithms to find the maximum likelihood estimate in these models.

As a result, a large number of alternative methods for estimating adjusted risk differences have been proposed, but these are often inflexible, prone to inefficiency or can produce invalid fitted risks. In this chapter, we remove the need for such approximate methods by describing a novel method for maximum likelihood estimation in additive binomial models that ensures that the parameter estimates are valid and that does not suffer the convergence issues associated with gradient-based approaches. The method

also permits flexible semi-parametric regression through the inclusion of isotonic regression functions.

The content in this chapter is reproduced from the published peer-reviewed journal article:

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* **80**: 184–196. DOI: 10.1016/j.csda.2014.06.019.

Minor editorial changes have been made to the article in order to maintain consistency across this thesis. The supplementary materials for the published article include comprehensive results of the simulations described in Section 3.4, which are reproduced here in Appendix 3.A.
R code for implementing the method described in this chapter was also provided as supplementary material for the article and is available online at http://dx.doi.org/10.1016/j.csda.2014.06.019. This has since been superseded by the R package `addreg`, which is available online at http://CRAN.R-project.org/package=addreg and its documentation is presented in Appendix B of this thesis.

Specific contribution of co-authors: I. C. Marschner assisted with conception of the method, and provided general supervision and feedback on research and writing. The candidate's contribution was at least 90% of the total effort required to produce the article.

Stable computational methods for additive binomial models with application to adjusted risk differences

Mark W. Donoghoe[1,2], Ian C. Marschner[1,2]

[1] Department of Statistics, Macquarie University, NSW 2109, Australia

[2] NHMRC Clinical Trials Centre, University of Sydney, NSW 2006, Australia

Abstract

Risk difference is an important measure of effect size in biostatistics, for both randomised and observational studies. The natural way to adjust risk differences for potential confounders is to use an additive binomial model, which is a binomial generalised linear model with an identity link function. However, implementations of the additive binomial model in commonly used statistical packages can fail to converge to the maximum likelihood estimate (MLE), necessitating the use of approximate methods involving misspecified or inflexible models. A novel computational method is proposed, which retains the additive binomial model but uses the multinomial–Poisson transformation to convert the problem into an equivalent additive Poisson fit. The method allows reliable computation of the MLE, as well as allowing for semi-parametric monotonic regression functions. The performance of the method is examined in simulations and it is used to analyse two datasets from clinical trials in acute myocardial infarction. Source code for implementing the method in R is provided as supplementary material.

Keywords: Additive binomial model · Multinomial–Poisson transformation · Risk difference · Semi-parametric regression

Reprinted from *Computational Statistics and Data Analysis* **80**, M. W. Donoghoe and I. C. Marschner, Stable computational methods for additive binomial models with application to adjusted risk differences: 184–196, Copyright (2014), with permission from Elsevier.

3.1 Introduction

In biostatistical applications, the risk of an event is the probability of the event occurring within a specific time frame. Risk difference is then the absolute difference in risk between two groups and is an important measure of effect size. For example, in randomised clinical trials, risk difference can be used to measure the magnitude of the treatment effect, while in observational studies it can be used to quantify the association between a risk factor and a disease event. Risk difference is important in practice because it is easier to interpret than the odds ratio and can present an alternative perspective to the relative risk.

As an important measure of effect size, the risk difference often needs to be adjusted for covariates. Analogous to logistic regression for estimating adjusted odds ratios, the natural model for estimating adjusted risk differences is a binomial generalised linear model (GLM) with identity link function, which we refer to as the additive binomial model.

The purpose of this paper is to address some common computational difficulties that arise with the additive binomial model for adjusted risk difference estimation. These difficulties arise from the requirement that the parameter space is constrained so that the linear probability model only produces probabilities in $[0, 1]$. This means that the model fitting is a constrained optimisation problem, and implementations of Fisher scoring or related procedures in popular statistics packages may fail to converge. Such numerical instability can occur even when the maximum likelihood estimate (MLE) is in the interior of the parameter space.

In light of these problems, there have been many proposals for estimating adjusted risk differences without using the additive binomial model. These include regression-based methods such as ordinary least squares or Poisson GLMs. However, with these methods the model is misspecified and fitted probabilities are not restricted to the $[0, 1]$ range. Alternative approaches can only provide the adjusted risk difference for a single binary comparison, and are essentially approximations to estimates from the additive binomial model. As demonstrated later in the paper, approximate methods for adjusted risk differences can have some undesirable properties, including loss of efficiency and violation of the parameter constraints.

In this paper we show that it is possible to retain the natural additive binomial model for adjusted risk difference estimation, without introducing numerical instability into the model-fitting process. We propose a computational method that uses a novel combination of two existing tools, the multinomial–Poisson transformation and a stable method for fitting additive Poisson models. A useful property of our approach is that it can be extended to allow semi-parametric adjustment, which is not available in other approaches.

We begin by specifying the additive binomial model that can be used to estimate adjusted risk differences, along with specification of the constrained parameter space and likelihood function. We then discuss how this can be recast into an equivalent additive Poisson estimation problem, using the multinomial–Poisson transformation. This allows application of stable computational methods for the additive Poisson model in order to fit the additive binomial model. Subsequently we present a range of simulation studies and analyses of two clinical trial datasets which demonstrate the advantages of our approach over competing methods for adjusted risk difference estimation. To facilitate practical implementation of this approach we have provided R code in the supplementary materials for this paper.

3.2 Method outline

3.2.1 Model definition

We assume that there are n independent observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, where each observation Y_i is associated with a vector of A categorical and B continuous covariates. The covariate vector for observation i is therefore $\boldsymbol{x}_i = (\boldsymbol{u}_i, \boldsymbol{v}_i) = (u_{i1}, \ldots, u_{iA}, v_{i1}, \ldots, v_{iB})$. Without loss of generality, we assume that $u_{ia} \in \{1, \ldots, k_a\}$ and $v_{ib} \in \mathbb{R}$, where k_a is the number of levels of categorical covariate a.

In a binomial GLM, Y_i is the number of events observed in a fixed number N_i of independent Bernoulli trials, where each trial has an event probability $p(\boldsymbol{x}_i, \boldsymbol{\theta})$ for some parameter vector $\boldsymbol{\theta}$. This event probability is referred to as the risk. With an

identity link function, the risk is assumed to have an additive structure

$$p(\boldsymbol{x}_i, \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta_b v_{ib}, \tag{3.1}$$

with $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_A, \boldsymbol{\beta})$, where $\boldsymbol{\alpha}_a = (\alpha_a(1), \ldots, \alpha_a(k_a))$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_B)$. Model (3.1) requires A identifiability constraints $\alpha_a(r_a) = 0$ for $a = 1, \ldots, A$, where r_a is the chosen reference level for categorical covariate a.

The risk difference for comparing two covariate combinations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is the difference in risks $p(\boldsymbol{x}_1, \boldsymbol{\theta}) - p(\boldsymbol{x}_2, \boldsymbol{\theta})$. Thus, the parameter $\alpha_a(u)$ represents the risk difference for the u^{th} level of the a^{th} categorical covariate versus the reference level r_a, adjusted for the $A - 1$ other categorical covariates and B continuous covariates in the model. Likewise, β_b represents the adjusted risk difference associated with a one-unit increase in the b^{th} continuous covariate.

3.2.2 Parameter space and likelihood function

Since the linear functions $p(\boldsymbol{x}, \boldsymbol{\theta})$ are probabilities, they must lie in the interval $[0, 1]$ for all $\boldsymbol{x}$ in the $(A + B)$-dimensional covariate space $\mathcal{X}$. We will define $\mathcal{X}$ as the space containing all possible combinations of the observed values of the covariates, that is $\mathcal{X} = \mathcal{U} \times \mathcal{V}$, where

$$\mathcal{U} = \prod_{a=1}^{A} \{1, \ldots, k_a\},$$

represents all possible combinations of the categorical covariates, and

$$\mathcal{V} = \prod_{b=1}^{B} [v_b^{(0)}, v_b^{(1)}],$$

is the B-dimensional Cartesian product of the observed ranges of the continuous covariates, with $v_b^{(0)} = \min_i\{v_{ib}\}$ and $v_b^{(1)} = \max_i\{v_{ib}\}$.

We wish to find the MLE $\hat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}$, subject to the constraint that $\hat{\boldsymbol{\theta}}$ lies in the parameter space

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} : 0 \leq p(\boldsymbol{x}, \boldsymbol{\theta}) \leq 1, \, \boldsymbol{x} \in \mathcal{X}\}. \tag{3.2}$$

The likelihood function for the additive binomial model, excluding a constant term, is

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{Y}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i, \boldsymbol{\theta})^{Y_i} (1 - p(\boldsymbol{x}_i, \boldsymbol{\theta}))^{N_i - Y_i}.$$

This model can, in principle, be fitted by any GLM software that fits identity-link binomial models. However, this approach is often numerically unstable due to the box constraints specified by (3.2), which can be difficult to handle with standard computational methods such as Fisher scoring. We therefore consider a more reliable approach that involves the novel combination of two existing tools described in the next two subsections.

3.2.3 Multinomial–Poisson transformation

The multinomial–Poisson (MP) transformation, described by Baker (1994), relates the likelihood for a multinomial model to that of a Poisson model. In general, the MP transformation applies to observations $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ij}, \ldots)$ from a multinomial distribution with $j \in J_i$, where J_i is any set of outcome categories for individual i. Here we describe and apply the MP transformation for the special case of $J_i = \{1, 2\}$ for all i, that is, the binomial distribution.

If Y_i are observations from a binomial distribution with N_i trials and event probability $p_i(\boldsymbol{\theta})$ for some parameter vector $\boldsymbol{\theta}$, we define the functions g_{i1} and g_{i2} such that

$$\frac{g_{i1}(\boldsymbol{\theta})}{G_i(\boldsymbol{\theta})} = p_i(\boldsymbol{\theta}) \qquad \text{and} \qquad \frac{g_{i2}(\boldsymbol{\theta})}{G_i(\boldsymbol{\theta})} = 1 - p_i(\boldsymbol{\theta}),$$

where $G_i(\boldsymbol{\theta}) = g_{i1}(\boldsymbol{\theta}) + g_{i2}(\boldsymbol{\theta})$.

The likelihood function for $\boldsymbol{\theta}$, excluding a multiplicative constant, is therefore

$$\mathcal{L}_{\mathrm{B}}(\boldsymbol{\theta}; \boldsymbol{Y}) = \prod_{i=1}^{n} \left(\frac{g_{i1}(\boldsymbol{\theta})}{G_i(\boldsymbol{\theta})} \right)^{Y_i} \left(\frac{g_{i2}(\boldsymbol{\theta})}{G_i(\boldsymbol{\theta})} \right)^{N_i - Y_i}.$$

Using dummy parameters $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)$, with all $\varphi_i > 0$, the MP transformation of $\mathcal{L}_{\mathrm{B}}$ is

$$\mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta}, \boldsymbol{\varphi}; \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{j=1}^{2} (\varphi_i g_{ij}(\boldsymbol{\theta}))^{Z_{ij}} \exp(-\varphi_i g_{ij}(\boldsymbol{\theta})), \tag{3.3}$$

where $Z_{i1} = Y_i$ and $Z_{i2} = N_i - Y_i$.

The MLE of $\boldsymbol{\varphi}$ for fixed $\boldsymbol{\theta}$ is $\hat{\varphi}_i(\boldsymbol{\theta}) = N_i/G_i(\boldsymbol{\theta})$, and substituting this back into (3.3) gives

$$\mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}(\boldsymbol{\theta}); \boldsymbol{Z}) \propto \mathcal{L}_{\mathrm{B}}(\boldsymbol{\theta}; \boldsymbol{Y}).$$

Thus, following the work of Richards (1961) on profile likelihoods, the MLE of $\boldsymbol{\theta}$ and the information matrix are identical for $\mathcal{L}_{\mathrm{B}}(\boldsymbol{\theta})$ and $\mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta}, \boldsymbol{\varphi})$. This means that the MLE for the binomial model may be found by maximising $\mathcal{L}_{\mathrm{P}}$, which takes the same form as the likelihood for a Poisson model with

$$Z_{ij} \sim \text{Poisson}(\varphi_i g_{ij}(\boldsymbol{\theta})). \tag{3.4}$$

The problem of finding the MLE for an additive binomial model (3.1) can thus be transformed into one of finding the MLE of a Poisson model (3.4) which involves both multiplicative ($\boldsymbol{\varphi}$) and additive ($\boldsymbol{\theta}$) components.

For model (3.1), the multiplicative component of (3.4) can be eliminated by defining

$$g_{i1}(\boldsymbol{\theta}) = N_i p(\boldsymbol{x}_i, \boldsymbol{\theta}) \quad \text{and} \quad g_{i2}(\boldsymbol{\theta}) = N_i(1 - p(\boldsymbol{x}_i, \boldsymbol{\theta})). \tag{3.5}$$

Then $G_i(\boldsymbol{\theta}) = N_i$, and the MLEs of the dummy parameters are $\hat{\varphi}_i = 1$ for all i, meaning that the problem reduces to one of finding the MLE of an additive Poisson model.

Note also that the parameter space for $\boldsymbol{\theta}$ which restricts the probabilities $p_i(\boldsymbol{\theta})$ to lie within $[0, 1]$ is the same as that which requires both $g_{i1}(\boldsymbol{\theta}) \geq 0$ and $g_{i2}(\boldsymbol{\theta}) \geq 0$ for all i. That is, the parameter constraints on the binomial probabilities are equivalent to non-negativity constraints on the Poisson means in (3.4).

3.2.4 Additive Poisson regression

The MP transformation converts an additive binomial fit into an additive Poisson fit. However, although fitting an additive Poisson model tends to be more numerically stable than fitting an additive binomial model, it can still be subject to instability in standard software. We therefore make use of the method presented by Marschner (2010) for additive Poisson models, which always provides reliable convergence to the MLE. As well as numerical stability, this method also has a number of other advantages.

The approach described by Marschner (2010) is a stable variant of the Expectation–Maximisation (EM) algorithm, and applies to any identity-link Poisson GLM. The computational method is an example of a combinatorial EM algorithm, which was presented in general terms by Marschner (2014). The main advantage of this approach is that it reliably accommodates the required non-negativity constraints on the Poisson means $g_{ij}(\boldsymbol{\theta})$ in (3.5). In addition, the method has some flexible features that enhance its usefulness. Firstly, while always accommodating the non-negativity constraints on the Poisson means, the method allows the model fitting to be conducted either with or without non-negativity constraints on the individual regression parameters $\boldsymbol{\theta}$. This is a useful feature that we make use of in implementing our method in Section 3.3.2. Secondly, the method can accommodate semi-parametric monotone regression functions, which allows semi-parametric adjustment of risk differences.

Next we describe in detail how the combination of these two basic methods, the MP transformation and stable additive Poisson regression, yields a reliable method for the additive binomial model that can be used for adjusted risk difference estimation.

3.3 Additive binomial regression

3.3.1 Linear covariates

We will begin by examining the case of a single continuous covariate v_i, with no other covariates in the model, so (3.1) reduces to

$$p(v_i, \boldsymbol{\theta}) = \alpha_0 + \beta v_i.$$

Without loss of generality we use a rescaled version of the continuous covariate

$$v_i^* = \frac{2v_i - (v^{(0)} + v^{(1)})}{v^{(1)} - v^{(0)}},$$

where $v^{(0)} = \min_i\{v_i\}$ and $v^{(1)} = \max_i\{v_i\}$, so that $v_i^* \in [-1, 1]$. Accordingly, we have a rescaled parameter vector $\boldsymbol{\theta}^* = (\alpha_0^*, \beta^*)$, such that $p(v, \boldsymbol{\theta}) = p(v^*, \boldsymbol{\theta}^*)$, using

$$\alpha_0^* = \alpha_0 + \frac{v^{(0)} + v^{(1)}}{2}\beta \qquad \text{and} \qquad \beta^* = \frac{v^{(1)} - v^{(0)}}{2}\beta. \tag{3.6}$$

The MP transformation is useful for additive binomial models because an additive model for $p(\cdot,\cdot)$ implies an additive model for $1-p(\cdot,\cdot)$. Thus, as in (3.5), we can define

$$g_{ij}(\boldsymbol{\theta}^*) = N_i g_j(v_i^*, \boldsymbol{\theta}^*) \qquad j = 1, 2,$$

where

$$\begin{aligned} g_1(v_i^*, \boldsymbol{\theta}^*) &= p(v_i^*, \boldsymbol{\theta}^*) = \alpha_0^* + \beta^* v_i^* \\ g_2(v_i^*, \boldsymbol{\theta}^*) &= 1 - p(v_i^*, \boldsymbol{\theta}^*) = (1-\alpha_0^*) + \beta^*(-v_i^*). \end{aligned}$$

This leads to a unified additive model

$$g_j(V_{ij}, \boldsymbol{\theta}^*) = \delta_j + \beta^* V_{ij}, \tag{3.7}$$

where $V_{ij} = (-1)^{j-1} v_i^*$ and $(\delta_1, \delta_2) = (\alpha_0^*, 1-\alpha_0^*)$. Note that $V_{ij} \in [-1, 1]$ for all i, j, ensuring that the covariate space is preserved in the unified model.

It follows from the MP transformation discussed in Section 3.2.3 that our problem of finding the MLE for the additive binomial model is equivalent to finding the MLE for an additive Poisson model with $2n$ observations, $\{(Z_{i1}, Z_{i2}), i = 1, \ldots, n\}$, where

$$Z_{ij} \sim \text{Poisson}(N_i g_j(V_{ij}, \boldsymbol{\theta}^*)), \tag{3.8}$$

with $Z_{i1} = Y_i$ and $Z_{i2} = N_i - Y_i$.

Model (3.8) requires the non-negativity constraints $g_j(V, \boldsymbol{\theta}^*) \geq 0$ for all $V \in [-1, 1]$, which ensures that $p(v^*, \boldsymbol{\theta}^*) \in [0, 1]$ for all $v^* \in [-1, 1]$. Fitting (3.8) subject to these constraints is achieved using the additive Poisson method of Marschner (2010) with one categorical covariate and one continuous covariate as specified by (3.7). The final step is then to transform $\boldsymbol{\theta}^*$ back onto its original scale using the relationships in (3.6).

Extension to $B > 1$ continuous covariates is straightforward. Each covariate is rescaled onto $[-1, 1]$, and the MLE for the rescaled additive binomial model is the same as the MLE for an additive Poisson model with one categorical covariate, B continuous covariates and $2n$ observations. This approach allows multiple linear regression models, which include non-linear polynomial models.

3.3.2 Categorical covariates

The approach used for continuous covariates does not apply directly to categorical covariates in an additive binomial model. However, a modification of this approach, again using the MP transformation, does allow incorporation of categorical covariates. We begin by considering the model with a single categorical covariate $u_i \in \{1, 2, \ldots, k\}$, so model (3.1) reduces to

$$p(u_i, \boldsymbol{\theta}) = \alpha_0 + \alpha_1(u_i). \tag{3.9}$$

Using the identifiability constraint $\alpha_1(1) = 0$, model (3.9) can be rewritten as a linear model

$$p(u_i, \boldsymbol{\theta}) = \alpha_0 + \sum_{b=2}^{k} \beta_b v_{ib}, \tag{3.10}$$

for an appropriately chosen parameterisation, β_b and v_{ib}. There are many possible parameterisations and a natural choice is

$$\beta_b = \alpha_1(b) \qquad \text{and} \qquad v_{ib} = \mathbf{1}\{u_i = b\}, \tag{3.11}$$

so that β_b is the contrast between level b and the reference level 1. The representation (3.10) would then seem to allow the categorical covariate model to be fitted using the methods described in Section 3.3.1 for linear covariates. In particular, note that the MP transformation described in Section 3.3.1 can again be applied, so that the additive binomial model (3.10) can be fitted using the equivalent additive Poisson model (3.8), with $k - 1$ linear covariates. However, there is a problem in that the procedure described in Section 3.3.1 will maximise the likelihood function over the parameter space that restricts the fitted event probabilities to be in $[0, 1]$ for all possible covariate combinations $(v_{i2}, \ldots, v_{ik})$ in which each v_{ib} is in $[0, 1]$. This is overly restrictive, because (3.11) does not allow more than one of the v_{ib} to equal to 1 for each i. Thus, the method of Section 3.3.1 applied to the linear covariate model (3.10) would impose additional constraints that would cause the likelihood function to be maximised over a smaller parameter space than is desired.

An alternative parameterisation is

$$\beta_b = \alpha_1(b) - \alpha_1(b-1) \qquad \text{and} \qquad v_{ib} = \mathbf{1}\{u_i \geq b\}, \tag{3.12}$$

so that the parameters β_b represent the increments between successive levels of the categorical covariate. This parameterisation has an analogous problem to that described above for parameterisation (3.11), so the method of Section 3.3.1 cannot be applied directly. However, the advantage of (3.12) is that it allows a simple modification that rectifies the problem.

As described in Section 3.2.4, Marschner (2010) presented a method for fitting the additive Poisson model, which can be applied subject to non-negativity constraints on the regression coefficients. When applied in the present context this method is an EM algorithm that imposes the constraints $\beta_b \geq 0$ for all $b = 2, \ldots, k$, or equivalently, $\alpha_1(1) \leq \alpha_1(2) \leq \cdots \leq \alpha_1(k)$. Although this imposes an undesired order restriction on the parameters, this constraint can be removed by repeatedly applying the order-restricted method after permuting the levels of the categorical covariate.

To see this, we first define the set $\mathcal{T}$, which consists of the $k!$ possible permutations of the levels of u_i. For each permutation $t \in \mathcal{T}$, there is a corresponding vector of permuted parameters $\big(\alpha_1^{(t)}(1), \alpha_1^{(t)}(2), \ldots, \alpha_1^{(t)}(k)\big)$. Application of the additive Poisson method with non-negativity constraints leads to maximisation of the likelihood over the space

$$\boldsymbol{\Theta}^{(t)} \subset \boldsymbol{\Theta} = \{\boldsymbol{\theta} : 0 \leq p(u, \boldsymbol{\theta}) \leq 1, \text{ for all } u = 1, \ldots, k\},$$

where $\boldsymbol{\Theta}^{(t)}$ is the subset of the parameter space $\boldsymbol{\Theta}$ that has $\alpha_1^{(t)}(1) \leq \alpha_1^{(t)}(2) \leq \cdots \leq \alpha_1^{(t)}(k)$. Since the parameter space $\boldsymbol{\Theta}$ may be partitioned into $k!$ such subsets corresponding to the $k!$ orderings $t \in \mathcal{T}$, it follows that $\boldsymbol{\Theta}$ is the union of these subsets, $\boldsymbol{\Theta} = \bigcup_t \boldsymbol{\Theta}^{(t)}$. Thus, having found the constrained maximum within each restricted parameter space $\boldsymbol{\Theta}^{(t)}$, the global maximum over $\boldsymbol{\Theta}$ will simply be the constrained maximum that achieves the highest likelihood. This procedure of cycling through all possible permutations of the categorical covariate levels, and applying an EM algorithm for each permutation, is an example of a combinatorial EM algorithm (Marschner, 2014).

For the model with $A > 1$ categorical covariates, where covariate a has k_a distinct levels, the same procedure is applicable, except that we must consider the Cartesian product of all possible permutations of each covariate. This leads to $K = \prod_{a=1}^{A} k_a!$ restricted parameter spaces that have to be searched for the MLE. In practice, if one of these spaces is found to have a stationary maximum, then it is the MLE and the algorithm may be halted. The same approach can be combined with the method

described in Section 3.3.1 for linear covariates, to fit the model with A categorical and B continuous covariates specified in (3.1). In this case the EM algorithm would be applied a maximum of $K = 2^{B+1} \prod_{a=1}^{A} k_a!$ times.

3.3.3 Data analysis example 1

The ASSENT-3 study (ASSENT-3 Investigators, 2001) was a clinical trial of 6095 patients with acute myocardial infarction (heart attack), randomly allocated to treatment regimens containing antithrombotic therapies. The primary treatment comparison of interest was between the group allocated to unfractionated heparin (UFH; $n = 2038$) and the group allocated to receive enoxaparin ($n = 2040$). The trial was designed as a non-inferiority study, with the non-inferiority margin being a 1% risk difference in favour of UFH for the composite endpoint of 30-day mortality and in-hospital reinfarction or ischaemia. As a brief numerical illustration we consider estimation of the risk difference between UFH and enoxaparin, adjusted for age.

For comparative purposes, we begin by investigating the form of the relationship between age, treatment and risk by fitting binomial GLMs with three different link functions: the logit link (adjusted odds ratio), the log link (adjusted relative risk) and the identity link (adjusted risk difference). We estimated the adjusted risk difference using the method presented above, as implemented in an R function called `addbin`. The resulting parameter estimates and standard errors (derived from the observed information matrix) were identical to those found using the `glm` function in R (R Core Team, 2013) and `PROC GENMOD` in SAS (SAS Institute Inc., 2008). For all three link functions, likelihood ratio tests for the inclusion of either a quadratic age relationship or an interaction between age and treatment were non-significant, and the parameter estimates and their approximate standard errors are shown in Table 3.1.

The deviances for the three alternative link functions are comparable, suggesting that any of these may be appropriate for modelling the risk of an event. This is an example of a scenario in which risk differences may be useful because they may be considered more interpretable than other measures, particularly odds ratios.

We will focus on the additive binomial model in order to obtain the treatment effect as a risk difference, adjusted for age. Whilst asymptotic normality would allow construction

TABLE 3.1: Comparison of adjusted effect measures in Example 1, based on GLMs with logit, log or identity link functions. Adjusted estimates are displayed for the odds ratio (OR), relative risk (RR) and risk difference (RD). Standard errors were estimated using the information matrix (SE_I) and bootstrap resampling (SE_B), and are shown on the log scale for the logit- and log-link models.

	Logit link			Log link			Identity link		
Parameter	OR	SE_I	SE_B	RR	SE_I	SE_B	RD	SE_I	SE_B
Treatment	0.70	0.093	0.093	0.74	0.080	0.080	−0.041	0.010	0.010
Age (per year)	1.03	0.004	0.004	1.03	0.003	0.003	0.003	0.0004	0.0003
Intercept	0.06	0.170	0.171	0.06	0.147	0.151	0.044	0.015	0.018
Scaled deviance		1.068			1.065			1.091	

of approximate confidence intervals, this assumption is questionable when the MLE is close to the boundary of the parameter space. We will demonstrate the stability of our algorithm by constructing non-parametric confidence intervals based on 1000 bootstrap samples.

The proposed method converged to the MLE in all 1000 bootstrap samples. In contrast, `PROC GENMOD` failed to converge in 214 samples. Figure 3.1 compares the distribution of the MLEs of each parameter, separated by whether or not `PROC GENMOD` reached convergence, demonstrating that confidence intervals obtained from only the converged samples would be biased.

The `glm` function available in R failed to converge in only one sample, however, this non-convergence was concerning because both `addbin` and `PROC GENMOD` converged to a point in the interior of the parameter space. Figure 3.2 shows the deviance achieved at each iteration of `glm`, demonstrating the potential instability of the algorithm implemented in R. This type of periodic non-convergence in R has been observed in other related contexts; see for example Marschner and Gillett (2012). In contrast, Figure 3.2 shows that `addbin` exhibited stable, albeit slow, convergence.

The additive model yields an estimate of 4.11% for the adjusted risk difference favouring the enoxaparin arm, with a one-sided 95% confidence interval from bootstrap resampling that extends to a 2.37% risk difference, still in favour of enoxaparin. This is well below the pre-specified 1% margin in favour of UFH, and so we can conclude that enoxaparin is not inferior (and in fact is superior) to UFH after adjusting for age. These adjusted results are consistent with the unadjusted results, which is not unexpected

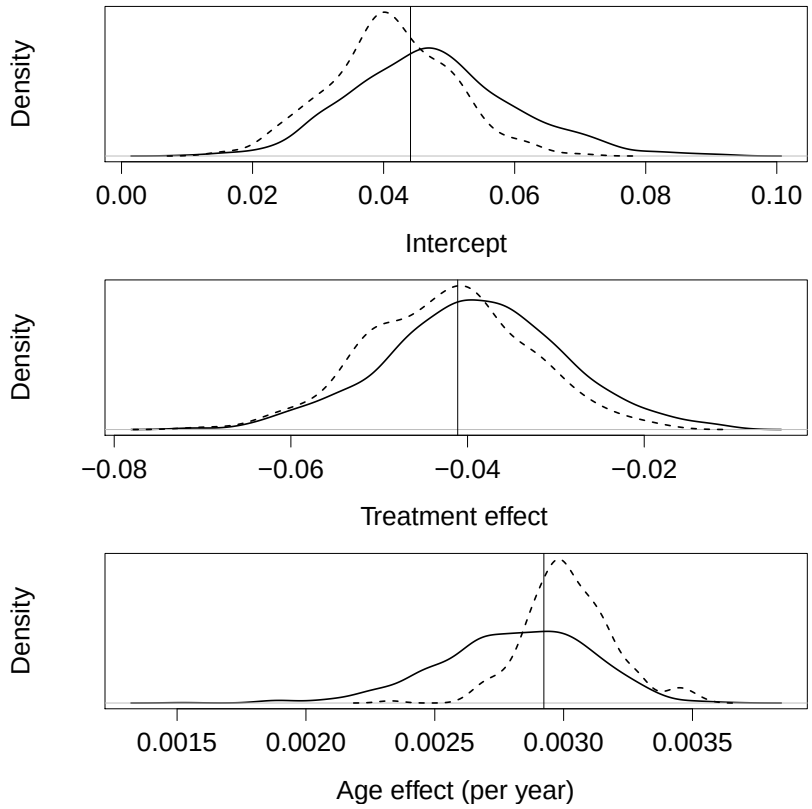


Figure 3.1: Smoothed density estimates of the MLE from 1000 bootstrap samples in Example 1, using `addbin` to compute the intercept (top), treatment (middle) and age (bottom) parameter estimates. Results are separated by whether `PROC GENMOD` converged (79%, solid line) or not (21%, dashed line). The vertical line shows the parameter estimate in the original data.

because age was balanced by randomisation. Nonetheless, the example does provide an initial numerical illustration of the method's performance.

3.4 Simulation study

For a more detailed evaluation of the performance of the MLE from an additive binomial GLM as an estimator of adjusted risk difference, we performed a number of simulation studies. In these simulations, the MLE was computed using the proposed method described in Section 3.3, as well as using the `glm` routine in R and `PROC GENMOD` in SAS. We empirically assessed the statistical properties of the proposed method, and compared it with various alternative non-MLE methods for calculating adjusted risk differences.

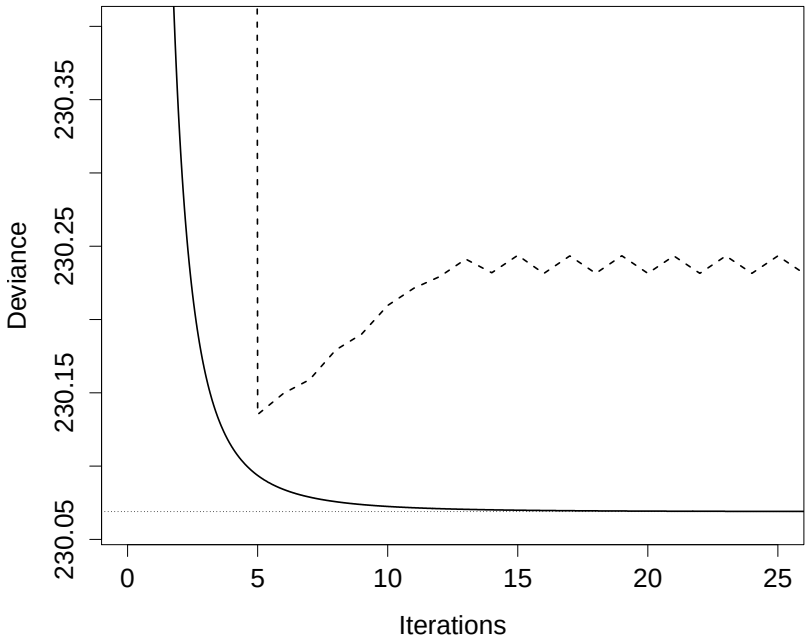


FIGURE 3.2: Iteration history for `glm` (dashed line) and `addbin` (solid line) in a bootstrap sample from Example 1. For the `addbin` iteration history, the number of iterations has been divided by 100. The dotted line denotes the optimal deviance.

3.4.1 Summary of alternative methods

Misspecified regression models

In cases where the additive binomial model fails to converge, alternative models have been suggested in which the distribution of the outcome variable is misspecified in order to estimate the model parameters. Using a Poisson GLM with an identity link was discussed by Spiegelman and Hertzmark (2005), and Cheung (2007) proposed modified least squares (MLS), where the binomial risk is represented as the expected value of a binary dependent variable, and ordinary least squares is used to find parameter estimates. In both cases, a robust variance estimator is used in calculating confidence intervals. The fitted risks from MLS are unrestricted, and the fitted Poisson means are only constrained to be non-negative, so both approaches can produce models with fitted risks outside $[0, 1]$.

Weighted mean methods

Other alternative methods only estimate the adjusted risk difference for a single binary comparison, rather than for a multivariable regression model. The first such methods

were based on data in the form of stratified 2×2 contingency tables, with the risk difference estimator being a weighted average of the unadjusted risk differences observed in each stratum. We examine weighting schemes defined by Cochran (1954) (Cochran–Mantel–Haenszel), Kleinbaum, Kupper, and Morgenstern (1982) (inverse variance), Rothman and Boice (1982) (null-weighted), Böhning and Sarol (2000), Greenland and Holland (1991), and Mehrotra and Railkar (2000) (minimum-risk). To avoid problems with zero cells, we follow Greenland and Robins (1985) and add $c = 0.5$ to each cell in calculating the inverse variance weights.

Other approximations

There exist other methods that are also restricted to a single binary comparison. Lee (1981) suggested fitting a logistic GLM and finding the average of the difference between the hypothetical fitted risks calculated as if all individuals had been in 'group 0' and those calculated as if all individuals had been in 'group 1'. Stijnen and Van Houwelingen (1993) proposed a pseudolikelihood approach for sparse stratified data, where the distribution of the response variable is misspecified as a standard normal distribution, such that nuisance parameters are removed from the likelihood and a consistent estimate for the adjusted risk difference can be found. Finally, Lunceford and Davidian (2004) proposed a number of estimators based on propensity scores, where the probability of group assignment must be modelled with respect to the adjustment variables. We examine the IPW2 estimator, later also derived by Ukoumunne et al. (2010), and the double-robustness estimator, which remains consistent if the model for the propensity scores is misspecified.

3.4.2 Simulation assumptions

We simulated samples of three different sizes, $n = 100$, 500 and 5000. Motivated by Example 1, the risk for individual i was determined by an additive model

$$p(\boldsymbol{x}_i, \boldsymbol{\theta}) = \alpha_0 + \alpha_1(u_i) + \beta_1 v_i,$$

where $u_i \in \{0, 1\}$ is the indicator for randomly-allocated treatment group (0 = control, 1 = intervention) and v_i is a continuous covariate for age, generated from a normal

distribution with mean 62.5 and variance 10^2, truncated to lie in the range $[40, 85]$. The parameter of interest is the adjusted risk difference between the treatment groups, $\alpha_1(1) - \alpha_1(0)$.

With the adjusted treatment effect taking values 0.05 and 0.15, and the gradient of age being 0.0015, 0.0030 or 0.0060 per year, we changed the value of α_0 to provide three different scenarios in which the properties of our method could be tested: an average risk of 0.5; a minimum risk of 0; and a maximum risk of 1.

For each sample size and set of parameter values, we produced 1000 simulations and estimated the parameters in an additive binomial model using the method described in Section 3.3, as implemented in the `addbin` routine. We estimated the bias of the risk difference parameter estimate, and calculated its sample variance.

We also calculated adjusted risk differences using each of the methods described in Section 3.4.1, and compared them to the MLE using the estimated mean squared error (MSE). For the misspecified regression models, which provide estimated risks for each individual, we counted the number of simulations in which all fitted risks were valid (within $[0, 1]$).

3.4.3 Results

The `addbin` routine found the MLE in all 1000 simulated samples for all 18 parameter combinations and all sample sizes, demonstrating its stability. The `glm` routine in R performed almost as well, but failed to converge in a small number of samples ($< 1\%$) with low or high risk ranges. `PROC GENMOD` in SAS converged to the MLE in over 99% of samples where the average risk was 0.5, but consistently failed to converge in around 50% of samples with risks close to 0 or 1, even with $n = 5000$.

Misspecified regression models

The MLE and the misspecified regression models all performed well for an average risk of 0.5, with the relative efficiency of the MLE being in the range 100–106%. Table 3.2 shows results for scenarios with 5% treatment effect and 0.6% age effect per year.

The additive Poisson model was slightly less efficient than the binomial MLE in terms of both the variance and MSE of its treatment effect estimate, but the difference was

usually less than 5%. The estimate from the least squares method (MLS) generally had a lower variance, though the gain in efficiency was less than 1%.

At the lower and upper ends of the risk range, MLS mostly produced estimates with slightly lower bias but much higher variance than the binomial MLE, and around half of these MLS models had fitted risks outside $[0, 1]$. This led to the MLE being more efficient, in the range 115–125% for this parameter combination. The Poisson model performed similarly to the MLE at low risks, where its non-negativity constraint on the fitted means was imposed. At high risks, the additive Poisson model often produced invalid fitted risks, and was less efficient than the MLE from the additive binomial, with relative efficiency around 125%.

The full range of results are presented in Section 3.A, Tables 3.A.1–3.A.6, where it is shown that the differences between methods may be smaller for other parameter combinations, albeit almost always in favour of the binomial MLE. Nonetheless, this example shows the potential for large differences in efficiency between the estimates.

Weighted mean methods

The weighted mean approaches generally performed best when the age covariate was split into 5 categories, and the results for $n = 100$ and $n = 5000$ are shown in Tables 3.A.7–3.A.12. When the average risk was 0.5, the performance of the alternative methods was generally similar to that of the MLE, with the exception of the inverse variance and Böhning–Sarol methods, which had inferior performance. Greenland and Holland's estimator performed particularly well in the interior of the parameter space for small n, having a small efficiency advantage (5%) compared to the MLE, but this was not consistently true close to the boundaries, where relative efficiency ranged from a 15% advantage to a 20% disadvantage in the parameter combinations we tested. Other estimators also generally suffered at least some loss of efficiency when risks were close to 0 or 1, with the exception of the null-weighted method in small samples, which often had a large bias but small variance, giving efficiency gains of up to 16% in some scenarios.

TABLE 3.2: Simulation results for adjusted treatment effect estimates from additive binomial, Poisson and modified least squares methods with a true treatment risk difference of 5% and age effect of 0.6% per year. Relative MSE is mean squared error relative to the binomial method, and "Valid" refers to the percentage of simulations with all fitted risks in $[0, 1]$.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE	Valid (%)
0.34–0.66	100	Binomial	7.25	0.0989	1	100
		Poisson	9.51	0.1017	1.058	99.4
		Least squares	6.74	0.0987	0.996	99.5
	500	Binomial	−1.72	0.0445	1	100
		Poisson	−1.15	0.0449	1.021	100
		Least squares	−1.57	0.0444	0.998	100
	5000	Binomial	−1.15	0.0141	1	100
		Poisson	−1.12	0.0143	1.022	100
		Least squares	−1.16	0.0141	0.999	100
0–0.32	100	Binomial	−6.95	0.0678	1	100
		Poisson	−6.68	0.0678	1.002	100
		Least squares	2.07	0.0745	1.205	47.8
	500	Binomial	−6.43	0.0285	1	100
		Poisson	−6.64	0.0287	1.009	100
		Least squares	−0.36	0.0307	1.140	49.5
	5000	Binomial	−1.51	0.0089	1	100
		Poisson	−1.42	0.0089	0.999	100
		Least squares	−0.48	0.0100	1.252	47.9
0.68–1	100	Binomial	−5.31	0.0669	1	100
		Poisson	6.32	0.0745	1.242	47.7
		Least squares	5.04	0.0719	1.156	48.7
	500	Binomial	−7.74	0.0300	1	100
		Poisson	−1.48	0.0338	1.248	49.4
		Least squares	−1.61	0.0329	1.188	48.4
	5000	Binomial	−1.12	0.0093	1	100
		Poisson	−0.22	0.0104	1.252	52.3
		Least squares	−0.13	0.0102	1.195	51.0

Other approximations

For the approaches that allow only a binary risk difference comparison to be made, we show the empirical properties relative to the binomial MLE in Tables 3.A.13–3.A.18. The estimates from the fitted logistic model and pseudolikelihood approach have almost identical properties, which are very similar to those of the binomial MLE when risks average 0.5. When risks were closer to 0 or 1, these estimators tended to be less efficient, resulting in 10–20% greater efficiency for the MLE in some scenarios. Both propensity score-based methods produced estimates with similar performance to the MLE in the interior of the parameter space. At the upper and lower boundaries, the propensity score methods generally had slightly lower bias but larger variance than the

MLE, resulting in efficiency losses of up to 25%.

3.4.4 Conclusions

Although there were some isolated scenarios in which alternative methods outperformed the additive binomial MLE for estimating adjusted risk difference, when viewed across the full range of scenarios, the MLE was the most efficient approach. Those estimators from misspecified regression models were generally less efficient than the correctly-specified binomial model, and also were not constrained to produce fitted risks inside $[0, 1]$. Other approximate methods required additional assumptions, and while they performed similarly to the MLE in the interior of the parameter space, they were substantially less efficient near the boundaries. The various weighted methods require that adjustment covariates be categorised, and only challenged the efficiency of the MLE in isolated scenarios.

3.5 Flexible monotonic regression

In some situations we may be confident of the direction of the effect of a continuous or ordered categorical covariate, but we may not want to restrict the relationship to be linear. To provide for more flexible modelling, we can include unspecified monotonic regression functions in our proposed method. This allows semi-parametric adjustment of risk differences, as well as exploration of an appropriate parametric form for the regression function.

We now include C monotonic covariates $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iC})$ in the model, where the contribution of w_{ic} to risk is determined by an unspecified non-decreasing function f_c. This leads to a semi-parametric extension of model (3.1):

$$p(\boldsymbol{x}_i, \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta_b v_{ib} + \sum_{c=1}^{C} f_c(w_{ic}). \tag{3.13}$$

The function f_c is only estimable at the unique observed values of w_{ic}, $z_c(0) < \cdots < z_c(l_c)$, and so for each monotonic covariate we introduce l_c parameters

$$\gamma_c(d) = f_c(z_c(d)) - f_c(z_c(d-1)), \qquad d = 1, \ldots, l_c,$$

with $\gamma_c(0) = f_c(z_c(0)) = 0$. These parameters represent the non-negative increments in risk between the observed covariate values. Model (3.13) can then be rewritten in the linear form

$$p(\boldsymbol{x}_i, \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta_b v_{ib} + \sum_{c=1}^{C} \sum_{d=1}^{l_c} \gamma_c(d) h_{icd},$$

where $h_{icd} = \mathbf{1}\{z_c(d) \leq w_{ic}\}$.

This form of the model means that for each monotonic covariate c we have l_c dummy linear covariates $\{h_{icd}; d = 1, \ldots, l_c\}$. These covariates can therefore be handled in the same manner as the linear covariates discussed in Section 3.3.1 after transformation to an additive Poisson model. The only difference with Section 3.3.1 is that the constraints $\gamma_c(d) \geq 0$ must be imposed to retain monotonicity of the regression function f_c. This can be handled straightforwardly by the additive Poisson method of Marschner (2010), which as discussed previously, allows such non-negativity constraints on the parameters.

3.5.1 Data analysis example 2

In Section 3.3.3, we showed an example in which different link functions produced similar fit, and our method allowed us to estimate an adjusted risk difference with bootstrapped confidence intervals. Here we demonstrate an example in which an additive binomial model provides a superior fit, but this is only apparent after identifying an appropriate functional form for a continuous covariate by first including it in the model as a semi-parametric monotonic covariate.

The ASSENT-2 study was a double-blind clinical trial to assess the safety and efficacy of tenecteplase versus alteplase in 16,949 patients with acute myocardial infarction (MI) treated within six hours (ASSENT-2 Investigators, 1999). The primary outcome was 30-day mortality after randomisation.

Marschner and Gillett (2012) analysed the ASSENT-2 data using a binomial GLM with a log link function, focusing on the age-specific relative risk of mortality, adjusting for MI severity, treatment delay and geographic region. Since mortality in the treatment arms was virtually identical, treatment was not included in the model.

We repeated the same analysis of the ASSENT-2 data, but this time with an additive binomial model, such that the parameters represent adjusted risk differences. With

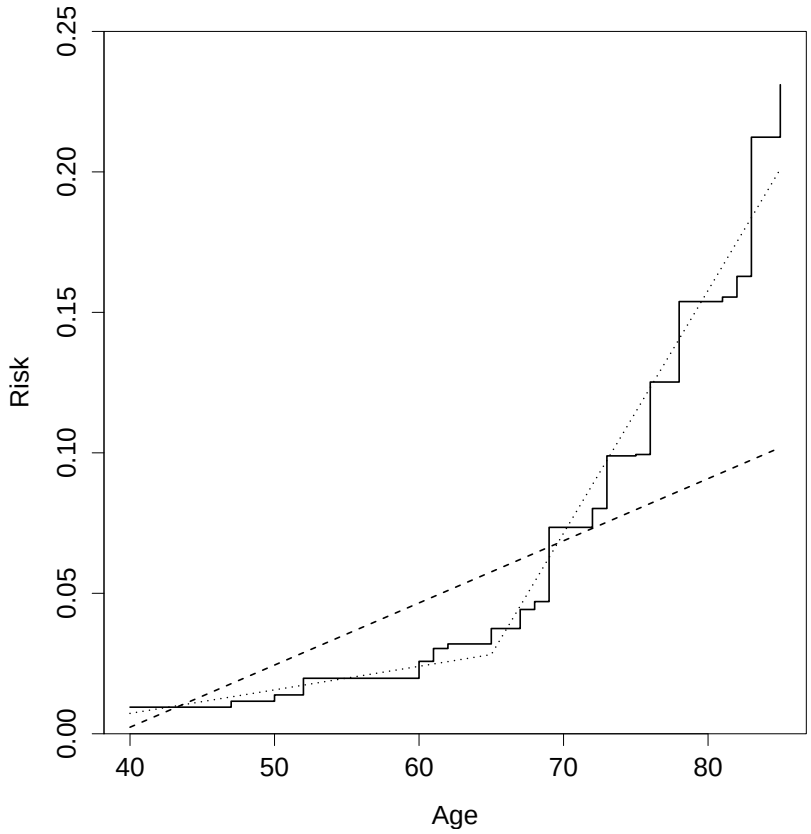


Figure 3.3: Age-specific risk of heart attack mortality in Example 2, using an additive risk model fitted with `addbin` and adjusted for event severity, treatment delay and region. The effect of age is specified using linear (dashed), piecewise linear (dotted) and semi-parametric monotonic (solid) regression functions. Fitted risks are presented for the Western region with low severity event and treatment delay < 2 hours.

age as a 3-level categorical variable (40–59, 60–75, 76–85 years), and adjusted for MI severity (Killip class I, II or III/IV), treatment delay (< 2, 2–4, > 4 hours) and region (Western countries, Latin America, or Eastern Europe), the residual deviance of the additive risk model fitted using `addbin` was 91.92 on 65 degrees of freedom compared to 149.32 for the relative risk model, indicating a superior fit. The fitted risks lie within $[0, 1]$, and both the `glm` function in R and `PROC GENMOD` in SAS successfully converged to the MLE in the main analysis. However, in 1000 bootstrap samples taken in order to estimate 95% confidence intervals, while both our method and the `glm` function converged in 100% of replications, `PROC GENMOD` failed to converge in 2.4% of samples. Since a scaled deviance of $91.92/65 = 1.41$ is not adequate, we further investigated the effect of increasing age on risk by entering it into the model as a 46-level covariate using a flexible monotonic function, and adjusting for the same categorical covariates as above. The adjusted age-specific risk from this model is plotted in Figure 3.3, shown

for the following covariate pattern: low severity event, < 2 hour delay and Western region. The monotonic model is compared to a model in which age is assumed to have a linear effect on risk, as well as a model with a piecewise linear effect of age, where the risk gradient changes at 65 years, as suggested by the shape of the monotonic function. The linear model is clearly inadequate, having a deviance of 926.13 on 744 degrees of freedom. The piecewise linear model, however, compares favourably to the monotonic model and has an adequate fit to the data with a deviance of 722.36 on 743 degrees of freedom.

The parameter values and their 95% confidence intervals (estimated using bootstrap resampling) for the piecewise linear model are shown in Table 3.3, compared to those from the model with age as a simple linear covariate. Under the piecewise linear model, each year of age up to 65 leads to a mortality risk difference of 0.08%, adjusted for MI severity, treatment delay and region. After 65 years of age, this rises to a risk difference of 0.9% per year. This illustrates that flexible monotonic regression, which is not available in standard implementations of the additive binomial GLM such as in R and SAS, can suggest a simpler parametric form for modelling risk. In constructing confidence intervals for the piecewise linear model, our method converged in all 1000 bootstrap samples, whereas the `glm` function in R failed to converge in one sample and `PROC GENMOD` in SAS failed to converge in over 20% of samples.

3.6 Discussion

We have described a computational method for estimating adjusted risk differences using the additive binomial model. The proposed approach is a novel combination of two existing methods: the MP transformation and a combinatorial EM algorithm for additive Poisson regression. This leads to a reliable procedure for computing the additive binomial MLE, which avoids the convergence problems inherent in standard GLM algorithms such as Fisher scoring.

Since the proposed method retains the natural additive binomial model, it avoids the need for other approximate methods for adjusted risk differences, which require us to misspecify the outcome distribution, make additional assumptions, or classify covariates into a one-dimensional list of strata. Furthermore, even when a standard GLM

Table 3.3: Risk differences (RD) and 95% confidence intervals (CI) from an additive risk model with a linear age term, and those from an additive risk model with a piecewise linear age term on the Example 2 data.

	Linear model		Piecewise linear model	
	RD	95% CI	RD	95% CI
Age (per year):				
40–85	0.0022	(0.0020, 0.0023)	—	—
40–65	—	—	0.0008	(0.0006, 0.0011)
65–85	—	—	0.0086	(0.0077, 0.0096)
Severity (Killip class):				
I	0	—	0	—
II	0.066	(0.050, 0.082)	0.061	(0.045, 0.076)
III/IV	0.281	(0.226, 0.335)	0.269	(0.216, 0.322)
Delay:				
< 2 hours	0	—	0	—
2–4 hours	−0.000	(−0.003, 0.003)	−0.001	(−0.006, 0.004)
> 4 hours	0.003	(−0.002, 0.010)	0.002	(−0.005, 0.009)
Region:				
Western	0	—	0	—
Latin America	−0.002	(−0.004, 0.001)	−0.002	(−0.008, 0.016)
Eastern Europe	0.032	(0.013, 0.052)	0.037	(0.018, 0.056)
Deviance	926.13 (744 df)		722.36 (743 df)	

algorithm does converge for an additive binomial model, the proposed method may still be advantageous for auxiliary analyses such as the bootstrap, which require convergence in many samples.

Standard algorithms such as Fisher scoring can sometimes be modified to increase their stability. One such approach is offered by the `glm2` package in R (Marschner, 2011), which uses a modified step-halving algorithm to ensure that the deviance will decrease at each iteration. This method converged to the MLE for the model in Example 1, and had a greater percentage of convergence in our simulations than the standard `glm` function, but still failed to converge in some samples. An alternative to standard GLM methods is a generic constrained optimisation algorithm applied to the additive binomial model. For example, Kovalchik et al. (2013) recently developed a method based on an adaptive barrier approach, which also includes the more general LEXPIT model. However, this too failed to converge to the MLE in some of our simulations. Importantly, our method has an advantage over all others in that it allows for the additional flexibility of unspecified monotonic regression functions. This allows semi-parametric adjustment of risk differences, and can also assist in identifying the functional form of

continuous covariates.

Adjusted risk differences have wide applicability in biostatistics, and this has led to the use of the additive binomial model in real applications (e.g. Grotvedt et al., 2008; Adelstein et al., 2011). From an individual perspective, risk difference is often a better effect measure than relative risk or odds ratio. Furthermore, from a population health perspective, risk difference is often more relevant than relative measures for assessing the benefit of a population intervention policy. One reason for this is that the reciprocal of the risk difference can be interpreted as the average number of individuals from the population that need to be treated with the intervention for a given time period to observe one fewer event within that time compared to the control, commonly referred to as the number needed to treat (Laupacis, Sackett, and Roberts, 1988).

In some datasets the additive binomial model may better characterise the simultaneous contribution of risk factors to an absolute change in risk, compared to a multiplicative model such as logistic regression. For example, the presence of an interaction between covariates on a multiplicative scale may disappear when their effects are considered on an additive scale, leading to a more parsimonious model for risk.

Our method for additive binomial models has been developed with adjusted risk differences in mind, but this model is also appropriate in many other situations. In epidemiology, adjusted prevalence differences from cross-sectional studies can be estimated using the additive binomial model. Linear probability models are also used in econometrics (Gujarati, 2003) and psychometrics (Maydeu-Olivares, 2005). This suggests that the proposed method may have broad applicability beyond our primary motivation of adjusted risk difference estimation.

Supplementary material

Supplementary material related to this article, including R code for implementing this method, can be found online at `http://dx.doi.org/10.1016/j.csda.2014.06.019`. Tables 3.A.1–3.A.18, referenced in Section 3.4.3 are reproduced in Appendix 3.A.

Acknowledgements

The authors thank the ASSENT-2 and ASSENT-3 investigators for providing data to the National Health and Medical Research Council Clinical Trials Centre. This research was supported by the Australian Research Council (DP110101254).

References

Adelstein, B.-A., T. A. Dobbins, C. A. Harris, I. C. Marschner, and R. L. Ward (2011). A systematic review and meta-analysis of KRAS status as the determinant of response to anti-EGFR antibodies and the impact of partner chemotherapy in metastatic colorectal cancer. *European Journal of Cancer* **47**(9): 1343–1354. DOI: 10.1016/j.ejca.2011.03.031.

ASSENT-2 Investigators (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *The Lancet* **354**(9180): 716–722. DOI: 10.1016/S0140-6736(99)07403-6.

ASSENT-3 Investigators (2001). Efficacy and safety of tenecteplase in combination with enoxaparin, abciximab, or unfractionated heparin: The ASSENT-3 randomised trial in acute myocardial infarction. *The Lancet* **358**(9282): 605–613. DOI: 10.1016/S0140-6736(01)05775-0.

Baker, S. G. (1994). The multinomial–Poisson transformation. *The Statistician* **43**(4): 495–504. DOI: 10.2307/2348134.

Böhning, D. and J. Sarol (2000). Estimating risk difference in multicenter studies under baseline-risk heterogeneity. *Biometrics* **56**(1): 304–308. DOI: 10.1111/j.0006-341X.2000.00304.x.

Cheung, Y. B. (2007). A modified least-squares regression approach to the estimation of risk difference. *American Journal of Epidemiology* **166**(11): 1337–1344. DOI: 10.1093/aje/kwm223.

Cochran, W. G. (1954). Some methods for strengthening the common Chi-squared tests. *Biometrics* **10**(4): 417–451. DOI: 10.2307/3001616.

Greenland, S. and P. Holland (1991). Estimating standardized risk differences from odds ratios. *Biometrics* **47**(1): 319–322. DOI: 10.2307/2532517.

Greenland, S. and J. M. Robins (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **41**(1): 55–68. DOI: 10.2307/2530643.

Grotvedt, L., H. Stigum, R. Hovengen, and S. Graff-Iversen (2008). Social differences in smoking and snuff use among Norwegian adolescents: A population based survey. *BMC Public Health* **8**: 322. DOI: 10.1186/1471-2458-8-322.

Gujarati, D. N. (2003). *Basic Econometrics*. 4th ed. New York: McGraw-Hill: p. 582.

Kleinbaum, D. G., L. L. Kupper, and H. Morgenstern (1982). *Epidemiologic Research: Principles and Quantitative Methods*. New York: Van Nostrand Reinhold: p. 359.

Kovalchik, S. A., R. Varadhan, B. Fetterman, N. E. Poitras, S. Wacholder, and H. A. Katki (2013). A general binomial regression model to estimate standardized risk differences from binary response data. *Statistics in Medicine* **32**(5): 808–821. DOI: 10.1002/sim.5553.

Laupacis, A., D. L. Sackett, and R. S. Roberts (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* **318**(26): 1728–1733. DOI: 10.1056/NEJM198806303182605.

Lee, J. (1981). Covariance adjustment of rates based on the multiple logistic regression model. *Journal of Chronic Diseases* **34**(8): 415–426. DOI: 10.1016/0021-9681(81)90040-0.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23**(19): 2937–2960. DOI: 10.1002/sim.1903.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**(3): 666–683. DOI: 10.1198/jcgs.2010.09127.

Marschner, I. C. (2011). glm2: Fitting generalized linear models with convergence problems. *The R Journal* **3**(2): 12–15.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* **24**(6): 921–940. DOI: 10.1007/s11222-013-9411-7.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics* **13**(1): 179–192. DOI: 10.1093/biostatistics/kxr030.

Maydeu-Olivares, A. (2005). "Linear item response theory, nonlinear item response theory, and factor analysis: A unified framework". *Contemporary Psychometrics*. Ed. by A. Maydeu-Olivares and J. J. McArdle. New Jersey: Lawrence Erlbaum Associates: pp. 73–100.

Mehrotra, D. V. and R. Railkar (2000). Minimum risk weights for comparing treatments in stratified binomial trials. *Statistics in Medicine* **19**(6): 811–825. DOI: 10.1002/(sici)1097-0258(20000330)19:6<811::aid-sim390>3.0.co;2-z.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **23**(2): 469–475.

Rothman, K. J. and J. D. Boice (1982). *Epidemiologic Analysis with a Programmable Calculator*. 2nd ed. Chestnut Hill, MA: Epidemiology Resources Inc.: p. 12.

SAS Institute Inc. (2008). *SAS/STAT Software, Version 9.2*. Cary, NC. URL: http://www.sas.com/.

Spiegelman, D. and E. Hertzmark (2005). Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* **162**(3): 199–200. DOI: 10.1093/aje/kwi188.

Stijnen, T. and H. C. Van Houwelingen (1993). Relative risk, risk difference and rate difference models for sparse stratified data: A pseudo likelihood approach. *Statistics in Medicine* **12**(24): 2285–2303. DOI: 10.1002/sim.4780122406.

Ukoumunne, O. C., E. Williamson, A. B. Forbes, M. C. Gulliford, and J. B. Carlin (2010). Confounder-adjusted estimates of the risk difference using propensity score-based weighting. *Statistics in Medicine* **29**(30): 3126–3136. DOI: 10.1002/Sim.3935.

Appendix

3.A Supplementary material

The following tables summarise the simulation results comparing the adjusted treatment effect estimates from the additive binomial to alternative methods. Tables 3.A.1–3.A.6 compare the additive binomial to the misspecified regression methods, Tables 3.A.7–3.A.12 compare the additive binomial to weighted methods, and Tables 3.A.13–3.A.18 compare the additive binomial to approximate methods. Relative MSE is mean squared error relative to the binomial method. For the regression methods, "Valid" is the percentage of simulations in which all fitted risks were within $[0, 1]$.

TABLE 3.A.1: Additive binomial versus regression methods, with 5% treatment effect and 0.15% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE	Valid (%)
0.44–0.56	100	Binomial MLE	2.95	0.0988	1	100
		Poisson	3.38	0.1002	1.027	99.7
		Least squares	3.13	0.0988	1.000	99.8
	500	Binomial MLE	0.82	0.0444	1	100
		Poisson	0.90	0.0446	1.010	100
		Least squares	0.82	0.0444	1.000	100
	5000	Binomial MLE	−0.20	0.0141	1	100
		Poisson	−0.17	0.0141	1.001	100
		Least squares	−0.20	0.0141	1.000	100
0–0.12	100	Binomial MLE	−8.60	0.0459	1	100
		Poisson	−8.17	0.0460	1.005	99.7
		Least squares	−3.85	0.0461	1.003	37.2
	500	Binomial MLE	−1.93	0.0205	1	100
		Poisson	−1.84	0.0205	1.002	100
		Least squares	0.20	0.0206	1.005	50.3
	5000	Binomial MLE	0.50	0.0065	1	100
		Poisson	0.52	0.0065	0.999	100
		Least squares	1.01	0.0068	1.082	49.4
0.88–1	100	Binomial MLE	−4.95	0.0471	1	100
		Poisson	1.52	0.0478	1.026	36.7
		Least squares	0.56	0.0472	1.001	36.0
	500	Binomial MLE	−2.72	0.0205	1	100
		Poisson	−0.20	0.0207	1.021	50.4
		Least squares	−0.37	0.0206	1.011	50.1
	5000	Binomial MLE	−0.60	0.0064	1	100
		Poisson	0.04	0.0067	1.092	49.0
		Least squares	0.02	0.0066	1.082	47.3

Table 3.A.2: Additive binomial versus regression methods, with 5% treatment effect and 0.3% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE	Valid (%)
0.41–0.59	100	Binomial MLE	9.54	0.1006	1	100
		Poisson	9.14	0.1023	1.035	99.9
		Least squares	9.14	0.1005	0.997	99.8
	500	Binomial MLE	0.98	0.0440	1	100
		Poisson	1.25	0.0443	1.012	100
		Least squares	0.94	0.0440	1.001	100
	5000	Binomial MLE	−0.83	0.0147	1	100
		Poisson	−0.83	0.0148	1.009	100
		Least squares	−0.83	0.0147	1.000	100
0–0.19	100	Binomial MLE	−3.33	0.0539	1	100
		Poisson	−3.19	0.0543	1.012	100
		Least squares	3.47	0.0568	1.107	45.9
	500	Binomial MLE	−4.60	0.0241	1	100
		Poisson	−4.57	0.0242	1.006	100
		Least squares	−0.93	0.0252	1.079	49.2
	5000	Binomial MLE	−1.68	0.0073	1	100
		Poisson	−1.66	0.0073	1.004	100
		Least squares	−0.68	0.0076	1.086	50.5
0.82–1	100	Binomial MLE	−5.61	0.0566	1	100
		Poisson	−1.04	0.0608	1.154	47.1
		Least squares	−1.56	0.0597	1.111	46.8
	500	Binomial MLE	−6.42	0.0246	1	100
		Poisson	−1.68	0.0260	1.102	50.8
		Least squares	−1.90	0.0256	1.074	49.9
	5000	Binomial MLE	−1.17	0.0076	1	100
		Poisson	−0.29	0.0083	1.192	51.7
		Least squares	−0.32	0.0083	1.166	50.2

TABLE 3.A.3: Additive binomial versus regression methods, with 5% treatment effect and 0.6% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE	Valid (%)
0.34–0.66	100	Binomial MLE	7.25	0.0989	1	100
		Poisson	9.51	0.1017	1.058	99.4
		Least squares	6.74	0.0987	0.996	99.5
	500	Binomial MLE	−1.72	0.0445	1	100
		Poisson	−1.15	0.0449	1.021	100
		Least squares	−1.57	0.0444	0.998	100
	5000	Binomial MLE	−1.15	0.0141	1	100
		Poisson	−1.12	0.0143	1.022	100
		Least squares	−1.16	0.0141	0.999	100
0–0.32	100	Binomial MLE	−6.95	0.0678	1	100
		Poisson	−6.68	0.0678	1.002	100
		Least squares	2.07	0.0745	1.205	47.8
	500	Binomial MLE	−6.43	0.0285	1	100
		Poisson	−6.64	0.0287	1.009	100
		Least squares	−0.36	0.0307	1.140	49.5
	5000	Binomial MLE	−1.51	0.0089	1	100
		Poisson	−1.42	0.0089	0.999	100
		Least squares	−0.48	0.0100	1.252	47.9
0.68–1	100	Binomial MLE	−5.31	0.0669	1	100
		Poisson	6.32	0.0745	1.242	47.7
		Least squares	5.04	0.0719	1.156	48.7
	500	Binomial MLE	−7.74	0.0300	1	100
		Poisson	−1.48	0.0338	1.248	49.4
		Least squares	−1.61	0.0329	1.188	48.4
	5000	Binomial MLE	−1.12	0.0093	1	100
		Poisson	−0.22	0.0104	1.252	52.3
		Least squares	−0.13	0.0102	1.195	51.0

Table 3.A.4: Additive binomial versus regression methods, with 15% treatment effect and 0.15% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE	Valid (%)
0.39–0.61	100	Binomial MLE	0.33	0.0977	1	100
		Poisson	1.24	0.0991	1.030	99.8
		Least squares	0.08	0.0974	0.994	99.9
	500	Binomial MLE	−1.13	0.0441	1	100
		Poisson	−0.94	0.0442	1.006	100
		Least squares	−1.14	0.0441	1.000	100
	5000	Binomial MLE	0.20	0.0140	1	100
		Poisson	0.24	0.0140	1.001	100
		Least squares	0.20	0.0140	1.000	100
0–0.22	100	Binomial MLE	−3.58	0.0616	1	100
		Poisson	−3.38	0.0617	1.003	100
		Least squares	−2.05	0.0610	0.976	30.8
	500	Binomial MLE	0.14	0.0274	1	100
		Poisson	0.21	0.0274	1.000	100
		Least squares	0.46	0.0274	0.999	46.6
	5000	Binomial MLE	−0.18	0.0085	1	100
		Poisson	−0.17	0.0085	1.000	100
		Least squares	−0.04	0.0085	1.014	48.8
0.78–1	100	Binomial MLE	−2.29	0.0589	1	100
		Poisson	0.11	0.0587	0.991	33.7
		Least squares	−0.89	0.0580	0.969	34.4
	500	Binomial MLE	−0.38	0.0276	1	100
		Poisson	0.19	0.0277	1.008	48.8
		Least squares	−0.01	0.0276	0.997	48.8
	5000	Binomial MLE	0.02	0.0086	1	100
		Poisson	0.13	0.0087	1.027	51.0
		Least squares	0.11	0.0087	1.020	48.4

TABLE 3.A.5: Additive binomial versus regression methods, with 15% treatment effect and 0.3% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE	Valid (%)
0.36–0.64	100	Binomial MLE	1.34	0.1073	1	100
		Poisson	2.55	0.1091	1.036	99.2
		Least squares	1.15	0.1070	0.995	99.9
	500	Binomial MLE	−0.46	0.0443	1	100
		Poisson	−0.31	0.0445	1.010	100
		Least squares	−0.48	0.0443	1.001	100
	5000	Binomial MLE	0.34	0.0136	1	100
		Poisson	0.34	0.0136	1.005	100
		Least squares	0.34	0.0136	1.000	100
0–0.29	100	Binomial MLE	−2.13	0.0688	1	100
		Poisson	−1.87	0.0691	1.008	100
		Least squares	0.28	0.0683	0.983	45.1
	500	Binomial MLE	−1.53	0.0307	1	100
		Poisson	−1.40	0.0307	0.998	100
		Least squares	−0.87	0.0314	1.037	48.5
	5000	Binomial MLE	−0.43	0.0095	1	100
		Poisson	−0.41	0.0095	1.000	100
		Least squares	−0.33	0.0096	1.027	53.1
0.72–1	100	Binomial MLE	−1.46	0.0683	1	100
		Poisson	0.99	0.0694	1.033	47.0
		Least squares	0.13	0.0682	0.997	46.9
	500	Binomial MLE	−0.56	0.0306	1	100
		Poisson	0.61	0.0310	1.024	49.9
		Least squares	0.37	0.0308	1.010	50.1
	5000	Binomial MLE	−0.31	0.0094	1	100
		Poisson	−0.11	0.0097	1.065	52.8
		Least squares	−0.13	0.0096	1.048	52.0

Table 3.A.6: Additive binomial versus regression methods, with 15% treatment effect and 0.6% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE	Valid (%)
0.29–0.71	100	Binomial MLE	−1.79	0.1038	1	100
		Poisson	−0.51	0.1054	1.030	99.2
		Least squares	−1.87	0.1036	0.997	99.1
	500	Binomial MLE	1.14	0.0443	1	100
		Poisson	1.42	0.0443	1.002	100
		Least squares	1.13	0.0443	0.999	100
	5000	Binomial MLE	−0.05	0.0144	1	100
		Poisson	−0.04	0.0145	1.014	100
		Least squares	−0.05	0.0144	1.000	100
0–0.42	100	Binomial MLE	−5.62	0.0799	1	100
		Poisson	−5.22	0.0806	1.015	100
		Least squares	−2.20	0.0807	1.009	51.1
	500	Binomial MLE	−2.27	0.0339	1	100
		Poisson	−2.18	0.0339	1.002	100
		Least squares	−0.87	0.0352	1.069	49.6
	5000	Binomial MLE	−0.14	0.0104	1	100
		Poisson	−0.11	0.0104	1.003	100
		Least squares	0.09	0.0109	1.097	51.2
0.58–1	100	Binomial MLE	−5.91	0.0806	1	100
		Poisson	−1.34	0.0849	1.097	54.0
		Least squares	−2.31	0.0820	1.024	54.2
	500	Binomial MLE	−3.44	0.0356	1	100
		Poisson	−1.50	0.0374	1.083	49.9
		Least squares	−1.72	0.0365	1.036	49.9
	5000	Binomial MLE	−0.61	0.0105	1	100
		Poisson	−0.28	0.0112	1.137	50.2
		Least squares	−0.30	0.0109	1.081	49.8

TABLE 3.A.7: Additive binomial versus weighted methods, with 5% treatment effect and 0.15% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.44–0.56	100	Binomial MLE	2.95	0.0988	1
		CMH	4.29	0.1008	1.041
		Böhning–Sarol	8.02	0.1097	1.234
		Inverse variance	9.05	0.1083	1.202
		Null-weighted	2.63	0.1000	1.023
		Greenland–Holland	0.59	0.0974	0.972
		Minimum risk	4.96	0.1017	1.060
	5000	Binomial MLE	−0.20	0.0141	1
		CMH	−0.19	0.0141	1.002
		Böhning–Sarol	−0.64	0.0147	1.088
		Inverse variance	−0.01	0.0141	1.007
		Null-weighted	−0.18	0.0141	1.003
		Greenland–Holland	−0.27	0.0141	1.001
		Minimum risk	−0.15	0.0141	1.003
0–0.12	100	Binomial MLE	−8.60	0.0459	1
		CMH	−4.03	0.0467	1.031
		Böhning–Sarol	−2.54	0.0506	1.209
		Inverse variance	−21.57	0.0415	0.865
		Null-weighted	−24.41	0.0405	0.844
		Greenland–Holland	−20.13	0.0419	0.874
		Minimum risk	−9.38	0.0450	0.963
	5000	Binomial MLE	0.50	0.0065	1
		CMH	1.01	0.0068	1.078
		Böhning–Sarol	1.33	0.0073	1.248
		Inverse variance	−0.69	0.0066	1.033
		Null-weighted	−0.82	0.0066	1.028
		Greenland–Holland	1.00	0.0068	1.081
		Minimum risk	0.73	0.0067	1.051
0.88–1	100	Binomial MLE	−4.95	0.0471	1
		CMH	0.51	0.0481	1.039
		Böhning–Sarol	2.18	0.0520	1.214
		Inverse variance	−17.52	0.0431	0.870
		Null-weighted	−20.27	0.0420	0.839
		Greenland–Holland	−16.68	0.0430	0.861
		Minimum risk	−4.96	0.0464	0.971
	5000	Binomial MLE	−0.60	0.0064	1
		CMH	0.01	0.0067	1.084
		Böhning–Sarol	0.02	0.0071	1.226
		Inverse variance	−1.24	0.0064	1.026
		Null-weighted	−1.36	0.0064	1.027
		Greenland–Holland	0.00	0.0067	1.086
		Minimum risk	−0.21	0.0066	1.058

Table 3.A.8: Additive binomial versus weighted methods, with 5% treatment effect and 0.3% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.41–0.59	100	Binomial MLE	9.54	0.1006	1
		CMH	7.36	0.1022	1.032
		Böhning–Sarol	11.34	0.1098	1.192
		Inverse variance	12.01	0.1120	1.240
		Null-weighted	5.73	0.1015	1.016
		Greenland–Holland	3.46	0.0983	0.954
		Minimum risk	7.77	0.1032	1.051
	5000	Binomial MLE	−0.83	0.0147	1
		CMH	−0.85	0.0147	1.000
		Böhning–Sarol	−0.60	0.0158	1.154
		Inverse variance	−0.69	0.0148	1.004
		Null-weighted	−0.85	0.0147	1.000
		Greenland–Holland	−0.93	0.0147	0.998
		Minimum risk	−0.82	0.0147	1.001
0–0.19	100	Binomial MLE	−3.33	0.0539	1
		CMH	4.87	0.0585	1.175
		Böhning–Sarol	4.03	0.0614	1.295
		Inverse variance	−8.66	0.0526	0.956
		Null-weighted	−12.17	0.0510	0.905
		Greenland–Holland	−1.11	0.0560	1.078
		Minimum risk	1.78	0.0569	1.114
	5000	Binomial MLE	−1.68	0.0073	1
		CMH	−0.71	0.0076	1.089
		Böhning–Sarol	−0.67	0.0082	1.250
		Inverse variance	−2.26	0.0075	1.090
		Null-weighted	−2.41	0.0075	1.093
		Greenland–Holland	−0.72	0.0076	1.091
		Minimum risk	−0.97	0.0075	1.060
0.82–1	100	Binomial MLE	−5.61	0.0566	1
		CMH	−1.02	0.0600	1.123
		Böhning–Sarol	1.10	0.0639	1.271
		Inverse variance	−12.94	0.0533	0.898
		Null-weighted	−16.23	0.0514	0.842
		Greenland–Holland	−6.77	0.0569	1.011
		Minimum risk	−3.68	0.0584	1.064
	5000	Binomial MLE	−1.17	0.0076	1
		CMH	−0.31	0.0083	1.168
		Böhning–Sarol	−0.44	0.0089	1.366
		Inverse variance	−1.20	0.0077	1.032
		Null-weighted	−1.35	0.0077	1.033
		Greenland–Holland	−0.33	0.0083	1.170
		Minimum risk	−0.45	0.0081	1.127

TABLE 3.A.9: Additive binomial versus weighted methods, with 5% treatment effect and 0.6% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.34–0.66	100	Binomial MLE	7.25	0.0989	1
		CMH	6.88	0.1001	1.023
		Böhning–Sarol	6.73	0.1069	1.167
		Inverse variance	12.95	0.1092	1.222
		Null-weighted	4.66	0.1003	1.027
		Greenland–Holland	2.87	0.0964	0.949
		Minimum risk	6.83	0.1013	1.048
	5000	Binomial MLE	−1.15	0.0141	1
		CMH	−1.17	0.0141	1.002
		Böhning–Sarol	−1.17	0.0151	1.144
		Inverse variance	−1.02	0.0141	1.004
		Null-weighted	−1.18	0.0141	1.002
		Greenland–Holland	−1.24	0.0141	1.001
		Minimum risk	−1.14	0.0141	1.002
0–0.32	100	Binomial MLE	−6.95	0.0678	1
		CMH	1.88	0.0765	1.272
		Böhning–Sarol	0.47	0.0821	1.462
		Inverse variance	−6.60	0.0717	1.118
		Null-weighted	−10.85	0.0680	1.012
		Greenland–Holland	−1.02	0.0744	1.203
		Minimum risk	0.41	0.0755	1.237
	5000	Binomial MLE	−1.51	0.0089	1
		CMH	−0.45	0.0100	1.258
		Böhning–Sarol	−0.18	0.0108	1.470
		Inverse variance	−1.24	0.0094	1.107
		Null-weighted	−1.39	0.0093	1.105
		Greenland–Holland	−0.47	0.0100	1.261
		Minimum risk	−0.58	0.0098	1.206
0.68–1	100	Binomial MLE	−5.31	0.0669	1
		CMH	4.18	0.0740	1.224
		Böhning–Sarol	5.45	0.0781	1.364
		Inverse variance	−0.76	0.0694	1.076
		Null-weighted	−6.03	0.0655	0.960
		Greenland–Holland	0.69	0.0718	1.153
		Minimum risk	3.25	0.0729	1.186
	5000	Binomial MLE	−1.12	0.0093	1
		CMH	−0.07	0.0102	1.196
		Böhning–Sarol	−0.03	0.0109	1.380
		Inverse variance	−0.11	0.0097	1.083
		Null-weighted	−0.25	0.0096	1.078
		Greenland–Holland	−0.12	0.0102	1.198
		Minimum risk	−0.04	0.0100	1.157

Table 3.A.10: Additive binomial versus weighted methods, with 15% treatment effect and 0.15% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.39–0.61	100	Binomial MLE	0.33	0.0977	1
		CMH	−0.14	0.0996	1.040
		Böhning–Sarol	−0.56	0.1046	1.147
		Inverse variance	7.02	0.1103	1.287
		Null-weighted	−0.64	0.0999	1.046
		Greenland–Holland	−3.91	0.0961	0.971
		Minimum risk	1.02	0.1012	1.074
	5000	Binomial MLE	0.20	0.0140	1
		CMH	0.21	0.0140	1.002
		Böhning–Sarol	0.32	0.0149	1.136
		Inverse variance	0.38	0.0140	1.006
		Null-weighted	0.21	0.0140	1.001
		Greenland–Holland	0.13	0.0140	1.000
		Minimum risk	0.24	0.0140	1.002
0–0.22	100	Binomial MLE	−3.58	0.0616	1
		CMH	−2.19	0.0615	0.990
		Böhning–Sarol	−1.80	0.0669	1.172
		Inverse variance	−12.84	0.0600	1.039
		Null-weighted	−16.39	0.0583	1.046
		Greenland–Holland	−11.92	0.0570	0.934
		Minimum risk	−4.48	0.0609	0.982
	5000	Binomial MLE	−0.18	0.0085	1
		CMH	−0.06	0.0086	1.017
		Böhning–Sarol	−0.02	0.0090	1.134
		Inverse variance	−0.58	0.0086	1.026
		Null-weighted	−0.70	0.0086	1.035
		Greenland–Holland	0.00	0.0086	1.020
		Minimum risk	−0.15	0.0085	1.014
0.78–1	100	Binomial MLE	−2.29	0.0589	1
		CMH	−0.85	0.0588	0.996
		Böhning–Sarol	−1.26	0.0632	1.150
		Inverse variance	−11.24	0.0590	1.082
		Null-weighted	−14.54	0.0579	1.102
		Greenland–Holland	−10.22	0.0554	0.949
		Minimum risk	−3.09	0.0587	0.996
	5000	Binomial MLE	0.02	0.0086	1
		CMH	0.12	0.0087	1.020
		Böhning–Sarol	0.20	0.0092	1.148
		Inverse variance	−0.28	0.0087	1.011
		Null-weighted	−0.41	0.0087	1.017
		Greenland–Holland	0.15	0.0087	1.022
		Minimum risk	0.05	0.0087	1.015

TABLE 3.A.11: Additive binomial versus weighted methods, with 15% treatment effect and 0.3% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.36–0.64	100	Binomial MLE	1.34	0.1073	1
		CMH	0.37	0.1086	1.024
		Böhning–Sarol	0.65	0.1153	1.154
		Inverse variance	6.55	0.1188	1.234
		Null-weighted	−0.92	0.1079	1.012
		Greenland–Holland	−3.34	0.1045	0.951
		Minimum risk	1.16	0.1097	1.045
	5000	Binomial MLE	0.34	0.0136	1
		CMH	0.34	0.0136	0.999
		Böhning–Sarol	0.33	0.0146	1.161
		Inverse variance	0.50	0.0136	1.004
		Null-weighted	0.34	0.0136	0.999
		Greenland–Holland	0.26	0.0136	0.997
		Minimum risk	0.37	0.0136	1.000
0–0.29	100	Binomial MLE	−2.13	0.0688	1
		CMH	0.32	0.0698	1.027
		Böhning–Sarol	−0.24	0.0728	1.117
		Inverse variance	−6.50	0.0696	1.041
		Null-weighted	−10.72	0.0678	1.022
		Greenland–Holland	−3.76	0.0673	0.961
		Minimum risk	−0.96	0.0702	1.041
	5000	Binomial MLE	−0.43	0.0095	1
		CMH	−0.33	0.0096	1.026
		Böhning–Sarol	−0.30	0.0101	1.137
		Inverse variance	−0.66	0.0095	1.010
		Null-weighted	−0.80	0.0095	1.014
		Greenland–Holland	−0.25	0.0096	1.030
		Minimum risk	−0.39	0.0096	1.016
0.72–1	100	Binomial MLE	−1.46	0.0683	1
		CMH	−0.23	0.0693	1.031
		Böhning–Sarol	0.75	0.0760	1.238
		Inverse variance	−6.94	0.0688	1.039
		Null-weighted	−11.17	0.0661	0.996
		Greenland–Holland	−4.09	0.0672	0.976
		Minimum risk	−1.70	0.0689	1.020
	5000	Binomial MLE	−0.31	0.0094	1
		CMH	−0.14	0.0096	1.052
		Böhning–Sarol	−0.20	0.0103	1.202
		Inverse variance	−0.42	0.0094	1.007
		Null-weighted	−0.55	0.0094	1.012
		Greenland–Holland	−0.08	0.0097	1.059
		Minimum risk	−0.18	0.0096	1.039

Table 3.A.12: Additive binomial versus weighted methods, with 15% treatment effect and 0.6% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.29–0.71	100	Binomial MLE	−1.79	0.1038	1
		CMH	−2.03	0.1054	1.032
		Böhning–Sarol	−2.55	0.1144	1.217
		Inverse variance	3.92	0.1120	1.167
		Null-weighted	−2.90	0.1043	1.011
		Greenland–Holland	−5.47	0.1020	0.972
		Minimum risk	−1.35	0.1060	1.044
	5000	Binomial MLE	−0.05	0.0144	1
		CMH	−0.05	0.0144	1.002
		Böhning–Sarol	−0.09	0.0153	1.118
		Inverse variance	0.10	0.0145	1.006
		Null-weighted	−0.07	0.0144	1.003
		Greenland–Holland	−0.13	0.0144	1.001
		Minimum risk	−0.03	0.0145	1.003
0–0.42	100	Binomial MLE	−5.62	0.0799	1
		CMH	−1.91	0.0823	1.050
		Böhning–Sarol	−1.72	0.0868	1.168
		Inverse variance	−4.81	0.0816	1.039
		Null-weighted	−9.90	0.0773	0.959
		Greenland–Holland	−4.74	0.0805	1.011
		Minimum risk	−2.41	0.0820	1.042
	5000	Binomial MLE	−0.14	0.0104	1
		CMH	0.10	0.0109	1.103
		Böhning–Sarol	0.16	0.0119	1.309
		Inverse variance	−0.13	0.0106	1.050
		Null-weighted	−0.28	0.0107	1.054
		Greenland–Holland	0.20	0.0109	1.112
		Minimum risk	0.07	0.0108	1.083
0.58–1	100	Binomial MLE	−5.91	0.0806	1
		CMH	−2.30	0.0830	1.049
		Böhning–Sarol	−1.55	0.0882	1.185
		Inverse variance	−5.18	0.0833	1.065
		Null-weighted	−9.98	0.0778	0.955
		Greenland–Holland	−5.34	0.0811	1.010
		Minimum risk	−2.88	0.0827	1.043
	5000	Binomial MLE	−0.61	0.0105	1
		CMH	−0.30	0.0109	1.083
		Böhning–Sarol	−0.28	0.0115	1.188
		Inverse variance	−0.49	0.0107	1.031
		Null-weighted	−0.63	0.0106	1.031
		Greenland–Holland	−0.23	0.0110	1.091
		Minimum risk	−0.33	0.0108	1.065

TABLE 3.A.13: Additive binomial versus approximate methods, with 5% treatment effect and 0.15% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.44–0.56	100	Binomial MLE	2.95	0.0988	1
		Fitted logistic	3.11	0.0988	1.000
		Pseudolikelihood	3.13	0.0988	1.000
		IPW2 prop. score	2.94	0.0987	0.997
		DR prop. score	2.99	0.0988	1.000
	5000	Binomial MLE	−0.20	0.0141	1
		Fitted logistic	−0.20	0.0141	1.000
		Pseudolikelihood	−0.20	0.0141	1.000
		IPW2 prop. score	−0.20	0.0141	1.000
		DR prop. score	−0.20	0.0141	1.000
0–0.12	100	Binomial MLE	−8.60	0.0459	1
		Fitted logistic	−3.71	0.0463	1.011
		Pseudolikelihood	−3.85	0.0461	1.003
		IPW2 prop. score	−3.98	0.0461	1.005
		DR prop. score	−2.49	0.0464	1.016
	5000	Binomial MLE	0.50	0.0065	1
		Fitted logistic	1.01	0.0068	1.083
		Pseudolikelihood	1.01	0.0068	1.082
		IPW2 prop. score	1.01	0.0068	1.082
		DR prop. score	1.01	0.0068	1.082
0.88–1	100	Binomial MLE	−4.95	0.0471	1
		Fitted logistic	0.86	0.0473	1.007
		Pseudolikelihood	0.56	0.0472	1.001
		IPW2 prop. score	0.58	0.0472	1.002
		DR prop. score	−0.77	0.0477	1.022
	5000	Binomial MLE	−0.60	0.0064	1
		Fitted logistic	0.02	0.0066	1.083
		Pseudolikelihood	0.02	0.0066	1.082
		IPW2 prop. score	0.02	0.0066	1.082
		DR prop. score	0.01	0.0066	1.082

Table 3.A.14: Additive binomial versus approximate methods, with 5% treatment effect and 0.3% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.41–0.59	100	Binomial MLE	9.54	0.1006	1
		Fitted logistic	9.07	0.1004	0.996
		Pseudolikelihood	9.14	0.1005	0.997
		IPW2 prop. score	8.99	0.1004	0.996
		DR prop. score	9.13	0.1005	0.997
	5000	Binomial MLE	−0.83	0.0147	1
		Fitted logistic	−0.83	0.0147	1.000
		Pseudolikelihood	−0.83	0.0147	1.000
		IPW2 prop. score	−0.83	0.0147	1.000
		DR prop. score	−0.82	0.0147	1.000
0–0.19	100	Binomial MLE	−3.33	0.0539	1
		Fitted logistic	4.06	0.0571	1.122
		Pseudolikelihood	3.47	0.0568	1.107
		IPW2 prop. score	3.50	0.0569	1.112
		DR prop. score	4.37	0.0569	1.113
	5000	Binomial MLE	−1.68	0.0073	1
		Fitted logistic	−0.67	0.0076	1.088
		Pseudolikelihood	−0.68	0.0076	1.086
		IPW2 prop. score	−0.68	0.0076	1.086
		DR prop. score	−0.68	0.0076	1.085
0.82–1	100	Binomial MLE	−5.61	0.0566	1
		Fitted logistic	−1.15	0.0598	1.114
		Pseudolikelihood	−1.56	0.0597	1.111
		IPW2 prop. score	−1.48	0.0597	1.111
		DR prop. score	−2.56	0.0600	1.124
	5000	Binomial MLE	−1.17	0.0076	1
		Fitted logistic	−0.31	0.0083	1.167
		Pseudolikelihood	−0.32	0.0083	1.166
		IPW2 prop. score	−0.32	0.0083	1.166
		DR prop. score	−0.31	0.0083	1.166

TABLE 3.A.15: Additive binomial versus approximate methods, with 5% treatment effect and 0.6% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.34–0.66	100	Binomial MLE	7.25	0.0989	1
		Fitted logistic	6.72	0.0987	0.995
		Pseudolikelihood	6.74	0.0987	0.996
		IPW2 prop. score	6.80	0.0988	0.998
		DR prop. score	6.46	0.0988	0.998
	5000	Binomial MLE	−1.15	0.0141	1
		Fitted logistic	−1.16	0.0141	0.999
		Pseudolikelihood	−1.16	0.0141	0.999
		IPW2 prop. score	−1.16	0.0141	0.999
		DR prop. score	−1.17	0.0141	0.999
0–0.32	100	Binomial MLE	−6.95	0.0678	1
		Fitted logistic	2.57	0.0748	1.214
		Pseudolikelihood	2.07	0.0745	1.205
		IPW2 prop. score	2.08	0.0745	1.206
		DR prop. score	2.61	0.0747	1.211
	5000	Binomial MLE	−1.51	0.0089	1
		Fitted logistic	−0.50	0.0100	1.253
		Pseudolikelihood	−0.48	0.0100	1.252
		IPW2 prop. score	−0.48	0.0100	1.252
		DR prop. score	−0.48	0.0100	1.251
0.68–1	100	Binomial MLE	−5.31	0.0669	1
		Fitted logistic	5.06	0.0721	1.162
		Pseudolikelihood	5.04	0.0719	1.156
		IPW2 prop. score	5.11	0.0719	1.158
		DR prop. score	5.14	0.0723	1.169
	5000	Binomial MLE	−1.12	0.0093	1
		Fitted logistic	−0.15	0.0101	1.192
		Pseudolikelihood	−0.13	0.0102	1.195
		IPW2 prop. score	−0.13	0.0102	1.195
		DR prop. score	−0.14	0.0102	1.194

Table 3.A.16: Additive binomial versus approximate methods, with 15% treatment effect and 0.15% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.39–0.61	100	Binomial MLE	0.33	0.0977	1
		Fitted logistic	0.06	0.0973	0.993
		Pseudolikelihood	0.08	0.0974	0.994
		IPW2 prop. score	0.05	0.0974	0.995
		DR prop. score	−0.06	0.0975	0.997
	5000	Binomial MLE	0.20	0.0140	1
		Fitted logistic	0.20	0.0140	1.000
		Pseudolikelihood	0.20	0.0140	1.000
		IPW2 prop. score	0.20	0.0140	1.000
		DR prop. score	0.20	0.0140	1.000
0–0.22	100	Binomial MLE	−3.58	0.0616	1
		Fitted logistic	−1.89	0.0611	0.977
		Pseudolikelihood	−2.05	0.0610	0.976
		IPW2 prop. score	−2.05	0.0610	0.977
		DR prop. score	−1.74	0.0613	0.983
	5000	Binomial MLE	−0.18	0.0085	1
		Fitted logistic	−0.04	0.0085	1.014
		Pseudolikelihood	−0.04	0.0085	1.014
		IPW2 prop. score	−0.04	0.0085	1.014
		DR prop. score	−0.04	0.0085	1.014
0.78–1	100	Binomial MLE	−2.29	0.0589	1
		Fitted logistic	−0.76	0.0580	0.968
		Pseudolikelihood	−0.89	0.0580	0.969
		IPW2 prop. score	−0.93	0.0581	0.970
		DR prop. score	−1.27	0.0584	0.982
	5000	Binomial MLE	0.02	0.0086	1
		Fitted logistic	0.12	0.0087	1.020
		Pseudolikelihood	0.11	0.0087	1.020
		IPW2 prop. score	0.11	0.0087	1.020
		DR prop. score	0.12	0.0087	1.021

TABLE 3.A.17: Additive binomial versus approximate methods, with 15% treatment effect and 0.3% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.36–0.64	100	Binomial MLE	1.34	0.1073	1
		Fitted logistic	1.14	0.1070	0.994
		Pseudolikelihood	1.15	0.1070	0.995
		IPW2 prop. score	1.10	0.1070	0.994
		DR prop. score	1.11	0.1069	0.993
	5000	Binomial MLE	0.34	0.0136	1
		Fitted logistic	0.34	0.0136	1.000
		Pseudolikelihood	0.34	0.0136	1.000
		IPW2 prop. score	0.34	0.0136	1.000
		DR prop. score	0.34	0.0136	1.000
0–0.29	100	Binomial MLE	−2.13	0.0688	1
		Fitted logistic	0.41	0.0684	0.985
		Pseudolikelihood	0.28	0.0683	0.983
		IPW2 prop. score	0.30	0.0684	0.987
		DR prop. score	0.57	0.0686	0.992
	5000	Binomial MLE	−0.43	0.0095	1
		Fitted logistic	−0.33	0.0096	1.027
		Pseudolikelihood	−0.33	0.0096	1.027
		IPW2 prop. score	−0.33	0.0096	1.027
		DR prop. score	−0.33	0.0096	1.027
0.72–1	100	Binomial MLE	−1.46	0.0683	1
		Fitted logistic	0.27	0.0682	0.997
		Pseudolikelihood	0.13	0.0682	0.997
		IPW2 prop. score	0.18	0.0682	0.997
		DR prop. score	0.06	0.0684	1.004
	5000	Binomial MLE	−0.31	0.0094	1
		Fitted logistic	−0.13	0.0096	1.048
		Pseudolikelihood	−0.13	0.0096	1.048
		IPW2 prop. score	−0.13	0.0096	1.048
		DR prop. score	−0.13	0.0096	1.048

Table 3.A.18: Additive binomial versus approximate methods, with 15% treatment effect and 0.6% age effect.

Risk range	n	Method	Bias (%)	Standard deviation	Relative MSE
0.29–0.71	100	Binomial MLE	−1.79	0.1038	1
		Fitted logistic	−1.92	0.1036	0.996
		Pseudolikelihood	−1.87	0.1036	0.997
		IPW2 prop. score	−1.86	0.1035	0.995
		DR prop. score	−1.78	0.1035	0.996
	5000	Binomial MLE	−0.05	0.0144	1
		Fitted logistic	−0.05	0.0144	1.000
		Pseudolikelihood	−0.05	0.0144	1.000
		IPW2 prop. score	−0.05	0.0144	1.000
		DR prop. score	−0.05	0.0144	1.000
0–0.42	100	Binomial MLE	−5.62	0.0799	1
		Fitted logistic	−2.14	0.0808	1.012
		Pseudolikelihood	−2.20	0.0807	1.009
		IPW2 prop. score	−2.21	0.0807	1.011
		DR prop. score	−2.15	0.0809	1.015
	5000	Binomial MLE	−0.14	0.0104	1
		Fitted logistic	0.09	0.0109	1.099
		Pseudolikelihood	0.09	0.0109	1.097
		IPW2 prop. score	0.09	0.0109	1.096
		DR prop. score	0.09	0.0109	1.096
0.58–1	100	Binomial MLE	−5.91	0.0806	1
		Fitted logistic	−2.22	0.0819	1.021
		Pseudolikelihood	−2.31	0.0820	1.024
		IPW2 prop. score	−2.41	0.0821	1.026
		DR prop. score	−2.51	0.0821	1.028
	5000	Binomial MLE	−0.61	0.0105	1
		Fitted logistic	−0.30	0.0109	1.081
		Pseudolikelihood	−0.30	0.0109	1.081
		IPW2 prop. score	−0.30	0.0109	1.081
		DR prop. score	−0.30	0.0109	1.081

4

Semi-parametric regression

In Chapter 3, we described a method for maximum likelihood estimation in additive binomial models which allows us to estimate adjusted risk differences. By using a combinatorial EM algorithm, the method avoids the convergence issues that can occur with the usual gradient-based approaches, and ensures that the estimates always remain within the parameter space of the model.

This method stands alongside the approaches described by Marschner (2010) and Marschner and Gillett (2012) that we outlined in Section 2.2.3, which use CEM algorithms to fit additive Poisson and log-link binomial models respectively, in order to estimate adjusted rate differences and relative risks.

However, these models are somewhat restrictive if we wish to use them to estimate or adjust for the effect of a continuous covariate on the risk or rate of events. With a fully parametric model, we must specify in advance the functional form of the relationship between the covariate and the risk.

Some level of flexibility is provided in each method through semi-parametric isotonic

regression functions. We showed with an example in Section 3.5.1 how this can be useful in identifying a parsimonious functional form for a continuous covariate. In that case, the shape of the isotonic regression curve, which uses a large number of degrees of freedom, suggested that a piecewise linear model might be appropriate, and this substantially improved the fit over a simple linear term.

In other situations, however, the shape of the isotonic curve may not suggest a simple transformation of the continuous covariate, meaning that we cannot find a parsimonious model with adequate fit. In many contexts, it is also more plausible that the effect of a continuous covariate on the risk or rate of an event is smooth rather than changing suddenly at a point, as does a piecewise or step function.

In this chapter, we describe an approach that extends the CEM algorithms for additive Poisson, additive binomial and log-binomial models to allow for the inclusion of smooth semi-parametric terms. We demonstrate its use in the ASSENT-2 data from Section 3.5.1, including a smooth term in place of the piecewise linear function in estimating adjusted risk differences and relative risks.

The content in this chapter is reproduced from the published peer-reviewed journal article:

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* **11**(1): 91–108. DOI: 10.1515/ijb-2014-0044.

Minor editorial changes have been made to the published version of the article in order to maintain consistency across this thesis. The R packages that implement the methods described in the article are available online at http://CRAN.R-project.org/package=logbin and http://CRAN.R-project.org/package=addreg, with their documentation presented in Appendices A and B of this thesis.

Specific contribution of co-authors: I. C. Marschner assisted with conception of the method, and provided general supervision and feedback on research and writing. The candidate's contribution was at least 90% of the total effort required to produce the article.

Flexible regression models for rate differences, risk differences and relative risks

Mark W. Donoghoe[1,2], Ian C. Marschner[1,2]

[1] Department of Statistics, Macquarie University, NSW 2109, Australia

[2] NHMRC Clinical Trials Centre, University of Sydney, NSW 2006, Australia

Abstract

Generalised additive models (GAMs) based on the binomial and Poisson distributions can be used to provide flexible semi-parametric modelling of binary and count outcomes. When used with the canonical link function, these GAMs provide semi-parametrically adjusted odds ratios and rate ratios. For adjustment of other effect measures, including rate differences, risk differences and relative risks, non-canonical link functions must be used together with a constrained parameter space. However, the algorithms used to fit these models typically rely on a form of the iteratively reweighted least squares algorithm, which can be numerically unstable when a constrained non-canonical model is used. We describe an application of a combinatorial EM algorithm to fit identity-link Poisson, identity-link binomial and log-link binomial GAMs in order to estimate semi-parametrically adjusted rate differences, risk differences and relative risks. Using smooth regression functions based on B-splines, the method provides stable convergence to the maximum likelihood estimates, and it ensures that the estimates always remain within the parameter space. It is also straightforward to apply a monotonicity constraint to the smooth regression functions. We illustrate the method using data from a clinical trial in heart attack patients.

Keywords: B-splines · Generalised additive models · Risk models · Semi-parametric regression

Reprinted from *International Journal of Biostatistics* **11**(1), M. W. Donoghoe and I. C. Marschner, Flexible regression models for rate differences, risk differences and relative risks: 91–108. The final publication is available at www.degruyter.com. DOI: 10.1515/ijb-2014-0044.

4.1 Introduction

Binary and count response data are commonly encountered in biostatistical settings, and generalised linear models (GLMs) are used to estimate the adjusted effects of several covariates on the risk or rate of these outcomes. The link function in the GLM determines the scale on which the effect measure is expressed. Using the canonical link with a binomial model for binary data gives adjusted odds ratios, and with a Poisson model for count data gives adjusted rate ratios (McCullagh and Nelder, 1989).

In biostatistics, other effect measures such as adjusted rate differences, risk differences or relative risks are often of interest, meaning that a non-canonical link function must be used. Under such models, constraints on the parameter space are required to ensure that fitted rates are non-negative and fitted risks lie within $[0, 1]$. However, the fitting procedures for GLMs in standard statistical software typically rely on a form of the iteratively reweighted least squares algorithm, which can be numerically unstable and fail to converge when parameter space constraints are present (Marschner, 2011). General step-size optimisation approaches can improve stability when estimates are close to the boundary of the parameter space, but these are not guaranteed to converge in all situations.

Marschner (2010), Donoghoe and Marschner (2014), and Marschner and Gillett (2012) have described stable methods for finding the maximum likelihood estimate (MLE) of Poisson GLMs with an identity link, and binomial GLMs with identity and log links. The methods allow estimation of adjusted rate differences, risk differences and relative risks respectively, avoiding convergence problems. All of these are applications of the combinatorial Expectation–Maximisation (CEM) algorithm presented by Marschner (2014).

With these models, however, the functional form of any continuous covariates must be specified. Generalised additive models (GAMs) are an extension of GLMs that allow for extra flexibility through the inclusion of semi-parametric terms (Hastie and Tibshirani, 1990; Wood, 2006). This can potentially lead to a better model fit, or help to identify a more parsimonious model for the outcome. Model-fitting with GAMs often uses similar algorithms as those used for GLMs, and hence is subject to similar convergence issues, particularly with non-standard link functions. In fact, in some GAM packages such as

`PROC GAM` in SAS (SAS Institute Inc., 2008), only the canonical link is permitted. In other packages, such as the R packages discussed in Section 4.6, non-canonical links are permitted but can be numerically unstable.

In this paper, we extend the existing CEM algorithms for GLMs to GAMs by the addition of smooth semi-parametric functions based on B-splines (de Boor, 1978; Eilers and Marx, 1996). We begin by defining the general GAM in Section 4.2, and in Section 4.3 we explore the properties of the B-spline basis functions. In Section 4.4 we explain the method for finding the MLE of each of these models, as well as how to apply an optional monotonicity constraint to the smooth functions. In Section 4.5, we demonstrate our methods by applying them to data from a clinical trial in heart attack patients. In Section 4.6, we summarise other popular methods for fitting GAMs and their performance in this dataset.

4.2 Model specification

Consider a sample of independent random variables $(Y_1, \ldots, Y_n)$ with either

$$Y_i \sim \text{Bin}(N_i, \lambda_i) \qquad \text{or} \qquad Y_i \sim \text{Poisson}(N_i \lambda_i),$$

so that $\mathbb{E}(Y_i) = N_i \lambda_i$ for some fixed known N_i. The quantity λ_i is interpreted as a standardised mean which will be a probability (risk) for the binomial model and a rate for the Poisson model.

In a GAM, the standardised mean λ_i is related to a linear combination of covariates through the link function

$$g(\lambda_i) = \Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i; \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta_b v_{ib} + \sum_{c=1}^{C} f_c(w_{ic}). \tag{4.1}$$

Here, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iA})$ are categorical covariates, where, without loss of generality, each u_{ia} takes a discrete value in $\{1, \ldots, k_a\}$. Linear continuous covariates are denoted by $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iB})$, with each v_{ib} allowed to take any value in the range $[v_b^{(0)}, v_b^{(1)}]$, where $v_b^{(0)} = \min_i\{v_{ib}\}$ and $v_b^{(1)} = \max_i\{v_{ib}\}$. The flexible part of the model is included through the unspecified non-parametric functions $f_1, \ldots, f_C$, which take as input the continuous covariates $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iC})$, and have domain $w_{ic} \in [w_c^{(0)}, w_c^{(1)}]$.

In order to impose a finite dimensional structure on the problem, the MLE $\hat{f}_c$ of the unknown f_c is restricted to be in the space defined by a specified set of basis functions $B_{c1}, \ldots, B_{cD_c}$. That is, each f_c can be expressed as

$$f_c(w) = \sum_{d=1}^{D_c} \gamma_{cd} B_{cd}(w), \tag{4.2}$$

a restriction we denote by $f_c \in \mathcal{B}_c$. Thus, the problem becomes one of finding the MLE of the parameter vector $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\alpha} = (\alpha_1(1), \ldots, \alpha_A(k_A))$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_B)$ and $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{cD_c})$.

One possible choice of basis functions is a sequence of indicator functions that serve as the increments between successive unique observed values of the covariate. That is, if $z_c(0) < \ldots < z_c(D_c)$ are the $D_c + 1$ ordered unique values of w_{ic}, we can define D_c basis functions

$$B_{cd}(w) = \mathbf{1}\{w \geq z_c(d)\}, \qquad d = 1, \ldots, D_c.$$

The resulting f_c is a step function, and each γ_{cd} represents the change in $g(\lambda)$ associated with an increase in the covariate from $z_c(d-1)$ to $z_c(d)$. If we constrain these increments to be strictly non-negative, this is the semi-parametric isotonic model described by Marschner (2010), Donoghoe and Marschner (2014), and Marschner and Gillett (2012).

Under the step function model, a large number of degrees of freedom are sacrificed in order to ensure that the semi-parametric function estimate fits closely to the observed data. Here we will instead focus on a smooth semi-parametric regression technique, in which the number of parameters that need to be estimated is reduced and f_c is smooth.

This could be achieved with any of a wide range of basis functions (Wood, 2006, pp. 146–167). We will use the polynomial B-splines, as they are highly flexible and their properties allow us to easily integrate them into the existing CEM algorithms, as discussed in the next section.

4.3 B-splines

4.3.1 Definition and properties

The B-splines are a series of basis functions for polynomial splines, defined by a grid of q_c fixed turning points $\xi_{c1} < \cdots < \xi_{cq_c}$, where $\xi_{c1} = w_c^{(0)}$ and $\xi_{cq_c} = w_c^{(1)}$, are positioned at the extremes of the domain. The grid is expanded to form a sequence $\boldsymbol{\tau}_c$ of knots that determine the degree and continuity of the resulting curve (Ramsay, 1988).

We will restrict our focus to the case of B-splines of order 3, where each basis function is made up of a series of quadratic curves between each pair of turning points, with adjacent curves constrained to have equal gradient at their boundaries. The sequence of $q_c + 4$ knots is defined such that there are three knots at the lower boundary, three knots at the upper boundary, and one at each of the internal turning points. That is:

$$\begin{aligned} \tau_{c1} = \tau_{c2} = \tau_{c3} &= \xi_{c1} \\ \tau_{cd} &= \xi_{c(d-2)} \quad \text{for} \quad d = 4, \ldots, q_c + 1 \\ \tau_{c(q_c+2)} = \tau_{c(q_c+3)} = \tau_{c(q_c+4)} &= \xi_{cq_c}. \end{aligned}$$

Given a knot sequence, the $D_c = q_c + 1$ B-splines can be calculated recursively (de Boor, 1978, pp. 128–135). We use $B_{cd}(w)$ to represent the d^{th} basis function $B_d(w; \boldsymbol{\tau}_c)$ on knot sequence $\boldsymbol{\tau}_c$, and $\mathcal{B}_c = \mathcal{B}(\boldsymbol{\tau}_c)$ to denote the associated function space for f_c.

The B-splines are normalised such that

$$\sum_{d=1}^{D_c} B_{cd}(w) = 1 \tag{4.3}$$

for all w, meaning that a constraint must be applied to the parameters for each c to ensure that they are identifiable. We do this by setting $\gamma_{ct_c} = 0$ for some choice of $t_c \in \{1, \ldots, D_c\}$.

Each basis function $B_{cd}(w)$ is positive for all $w \in (\tau_{cd}, \tau_{c(d+3)})$ and zero elsewhere, taking its maximum value for some $w \in (\tau_{c(d+1)}, \tau_{c(d+2)})$. This means that each parameter γ_{cd} has only local influence on the smooth function f_c.

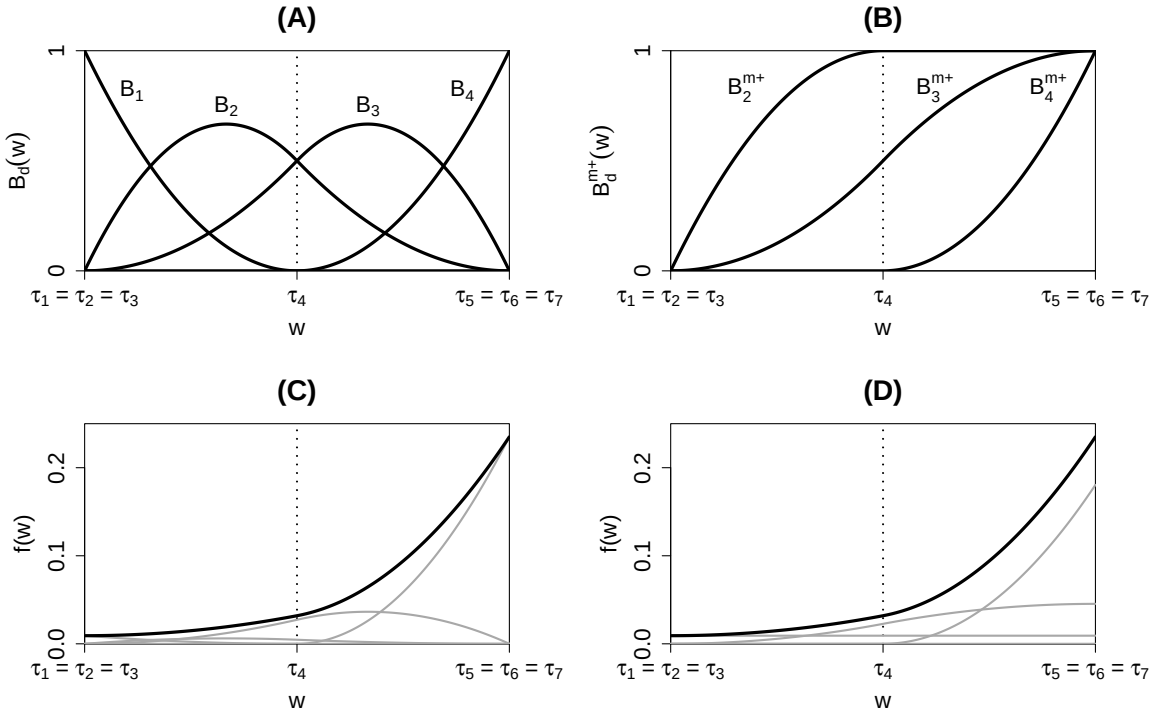


FIGURE 4.1: (A) B-spline basis functions and (C) example resulting curve with one internal knot, and (B) the equivalent monotonic B-spline basis functions and (D) resulting curve based on the same coefficients.

As an illustration, we present a graphical representation in Figure 4.1. The four B-spline basis functions of order 3 with a single internal turning point, and hence 7 knots, are shown in Figure 4.1(A). Figure 4.1(C) shows the curve f that results from taking a particular linear combination of these basis functions

$$f(w) = 0.009B_1(w) + 0.009B_2(w) + 0.055B_3(w) + 0.235B_4(w), \tag{4.4}$$

which can be equivalently expressed as

$$f(w) = 0.009 + 0.000B_2(w) + 0.045B_3(w) + 0.226B_4(w)$$

by using (4.3). Figures 4.1(B) and (D) illustrate the use of B-splines to restrict f to be monotonic, and will be discussed in Section 4.3.3.

4.3.2 Model constraints

The models considered in this paper have constraints on their parameter spaces due to the restricted range of the response variable: rates must be non-negative and risks must lie in $[0, 1]$. These constraints are often the source of instability in standard fitting algorithms, particularly when the MLE is close to the boundary of the parameter space (Marschner, 2011).

The CEM algorithm is ideal for fitting these models, as it applies the parameter space constraints while guaranteeing stable convergence to the MLE. A crucial step in the definition of a CEM algorithm is that the parameter space is partitioned into a sequence of subspaces, each of which corresponds to a particular set of constraints on individual parameters. The properties of the B-splines as discussed in Section 4.3.1 allow us to extend the existing methods to include these semi-parametric functions.

As an example, we restrict our attention to $f_c \in \mathcal{B}_c^+$, the space of strictly non-negative curves in $\mathcal{B}_c$. This can be done without loss of generality due to (4.3), such that the range of f_c is determined by the intercept α_0, and its shape by the B-spline coefficients. The function space $\mathcal{B}_c^+$ can be partitioned into subspaces defined by the index of the smallest coefficient. That is, if we define

$$\mathcal{B}_c^+(t_c) = \left\{ f \in \mathcal{B}_c^+ : \gamma_{ct_c} = \min_d \{\gamma_{cd}\} \right\}, \tag{4.5}$$

it is easy to see that

$$\mathcal{B}_c^+ = \bigcup_{t_c=1}^{D_c} \mathcal{B}_c^+(t_c). \tag{4.6}$$

For a particular identifiability constraint $\gamma_{ct_c} = 0$, if the remaining coefficients are restricted to be non-negative, the resulting f_c will be a strictly non-negative curve. Furthermore, γ_{ct_c} will be the smallest of the coefficients; that is, $f_c \in \mathcal{B}_c^+(t_c)$.

One characterisation of the constrained function space is that any curve $f_c(w) \in \mathcal{B}_c^+(t_c)$ will have a local minimum for $w \in [\tau_{c(t_c+1)}, \tau_{c(t_c+2)}]$. The special case of $t_c = 1$ corresponds to the family of non-negative curves that take their minimum at the lower limit of the domain, $w_c^{(0)}$, and likewise $t_c = D_c$ will constrain the curves to take their minimum at $w_c^{(1)}$. This is demonstrated with an example in Section 4.4.2.

An analogous result can be achieved by fixing $\gamma_{ct_c} = 0$ and restricting the remaining

parameters to be non-positive. Then $f_c(w)$ will be a non-positive curve with γ_{ct_c} being the largest coefficient, a function space we denote by $\mathcal{B}_c^-(t_c)$. Similarly to the non-negative case discussed above, $\mathcal{B}_c^-(t_c)$ can be characterised by its restriction on the location of local maxima.

A useful property of the existing CEM algorithms is that they can easily accommodate non-negativity or non-positivity constraints on individual coefficients. By applying such constraints, we can find the MLE of the GAM under the restriction that each $f_c \in \mathcal{B}_c^+(t_c)$ or $\mathcal{B}_c^-(t_c)$, and by repeating this process for all possible choices of the identifiability constraint, we can find the overall MLE. This process is explained in more detail for the specific models of interest in Section 4.4.

4.3.3 Monotonic B-splines

In some contexts, it may be sensible to restrict $f_c(w)$ to increase with increasing w. A sufficient condition for f_c to be monotonically non-decreasing is that the coefficients of the B-splines are themselves strictly non-decreasing (Leitenstorfer and Tutz, 2007), that is, $\gamma_{c1} \leq \cdots \leq \gamma_{cD_c}$.

Assuming non-negative coefficients, the identifiability constraint $\gamma_{c1} = 0$ ensures that $f_c(w_c^{(0)}) = 0$ is the minimum of the smooth function. We can then introduce $D_c - 1$ new parameters that represent the increments between successive B-spline coefficients:

$$\delta_{cd} = \gamma_{cd} - \gamma_{c(d-1)},$$

for $d = 2, \ldots, D_c$. Restricting these increments to be non-negative will then ensure that the original coefficients are monotonically non-decreasing, as desired.

This can be achieved in the same manner as the unrestricted case by re-expressing (4.2) as

$$\begin{aligned} f_c(w) &= \sum_{d=2}^{D_c} \left[\sum_{e=2}^{d} \delta_{ce} \right] B_{cd}(w) \\ &= \sum_{d=2}^{D_c} \delta_{cd} \left[\sum_{e=d}^{D_c} B_{ce}(w) \right] \\ &= \sum_{d=2}^{D_c} \delta_{cd} B_{cd}^{m+}(w). \end{aligned}$$

Here, $\{B_{cd}^{m+}, d = 2, \ldots, D_c\}$ denotes a new series of non-negative monotonic basis functions, which we will refer to as monotonic B-splines. In fact, for the order 3 B-splines used here, these new basis functions are equivalent to the piecewise quadratic integrated splines (I-splines) introduced by Ramsay (1988) and used by Tutz and Leitenstorfer (2007) for generalised smooth monotonic regression.

Continuing with the illustration provided in Section 4.3.1, the monotonic B-spline basis functions associated with the B-spline bases in Figure 4.1(A) are shown in Figure 4.1(B). Because the B-spline coefficients in (4.4) are monotonically non-decreasing, the resulting curve in Figure 4.1(C) can be expressed in terms of the monotonic B-splines

$$f(w) = 0.009 + 0.000B_2^{m+}(w) + 0.045B_3^{m+}(w) + 0.181B_4^{m+}(w),$$

and this is demonstrated in Figure 4.1(D), where the intercept term is shown as a horizontal line.

For the alternate case in which we wish for the curve f_c to be non-positive and monotonically non-decreasing with its maximum value at $w_c^{(1)}$, a similar process applies. We fix $\gamma_{cD_c} = 0$, and define $D_c - 1$ new parameters, $\delta_{cd} = \gamma_{cd} - \gamma_{c(d+1)}$ for $d = 1, \ldots, D_c - 1$. The associated monotonic basis functions are defined as

$$B_{cd}^{m-}(w) = \sum_{e=1}^{d} B_{ce}(w), \qquad d = 1, \ldots, D_c - 1, \tag{4.7}$$

and their coefficients δ_{cd} are constrained to be non-positive.

4.3.4 Knot selection

The number and placement of the turning points will influence the shape of the resulting function by determining the space in which our estimated f_c is constrained to lie. Ideally, external information would be used to determine the turning points; however, in many applications this will not be available and we are forced to depend on our data to guide this decision.

In some situations, such as OLS, the positioning of the turning points is crucially important (Ruppert, Wand, and Carroll, 2003), and a large number of knot selection

methods have been derived (e.g. de Boor, 1978, pp. 174–196, Friedman and Silverman, 1989). Many of these methods can be integrated into the approach we describe, but they often lead to a large increase in the computational burden. In a general setting, Ramsay (1988) noted that the shape of the resulting estimate is not particularly sensitive to knot placement.

We apply an adaptation of cardinal splines discussed by Hastie and Tibshirani (1990, p. 24), placing the $q_c - 2$ internal turning points at evenly spaced quantiles of the observed covariate values w_{ic}. This is often the standard approach used with regression splines (e.g. Ruppert, 2002). Alternatively, our implementation in R discussed in Section 4.6 optionally allows the user to specify their own list of knots, so other knot-placement regimes could be used, such as equally spaced knots, nested knot structures, or the approach proposed by Yao and Lee (2008), in which knots are placed at the local minima and maxima of an initial estimate of the smooth curve, fitted using a basic knot structure.

Of greater importance is the choice of the number of turning points. With too few, we may fail to detect important features of the relationship, but with too many, we are at risk of over-fitting. One common way to resolve this trade-off is to use a sufficiently large number of turning points to broaden the function space for f_c, but add a penalty term to the likelihood function such that spurious fluctuations in the smooth function are avoided (Green and Silverman, 1994).

However, with a penalty term, the CEM algorithm central to these stable methods cannot be used directly. We discuss this further, and propose a possible solution in Section 4.7. In general, we will use the Akaike information criterion $AIC = 2J - 2\ell$ (Akaike, 1974) to choose between models with different numbers of knots, where ℓ is the log-likelihood of the fitted model. This similarly includes a penalty for model complexity, and in the case of unpenalised maximum likelihood estimation, the effective degrees of freedom J is simply the number of estimated parameters in the model (Wood, 2006, pp. 170–171). The AIC_c is a bias-corrected version of the AIC for small samples (Burnham and Anderson, 2002, p. 66), which becomes virtually identical as n increases. Figure 4.4 illustrates the use of the AIC in determining the optimal number of knots for the example in Section 4.5.

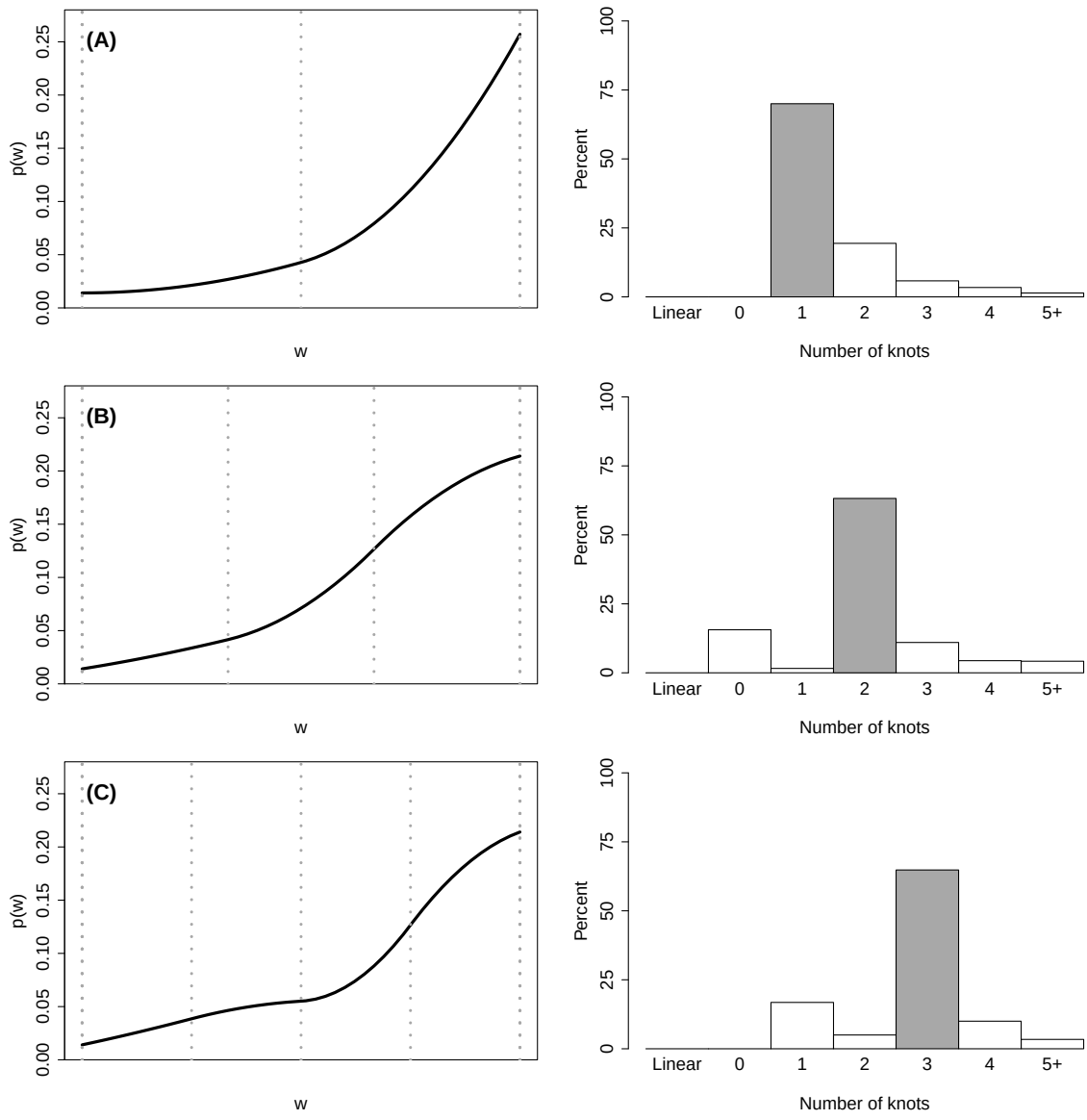


FIGURE 4.2: Results from 500 simulations of the performance of the AIC for selecting the optimal number of knots. The panels on the left-hand side show the true risk function $p(w)$ with (A) 1, (B) 2 and (C) 3 internal knots. The panels on the right-hand side show a histogram of the model selected by AIC in each simulation, with the bar corresponding to the true model shaded.

As a criterion for choosing the optimal smoothing parameter in a wide range of scenarios, the AIC and AIC_c have been shown to be generally superior to other classical approaches such as BIC and GCV in the context of general non-parametric regression by Hurvich, Simonoff, and Tsai (1998), semi-parametric Cox regression by Malloy, Spiegelman, and Eisen (2009) and likelihood-based boosting by Leitenstorfer and Tutz (2007).

Figure 4.2 shows the results from a simulation study in which we examined the performance of the AIC value in selecting the optimal number of knots. We simulated 500 datasets of 750 binomial observations, each with true risk functions that were based on B-splines with 1, 2 or 3 internal knots. For each dataset, we found the MLE corresponding to a binomial model in which the continuous variable was included as a linear term or as a B-spline with between 0 to 5 internal knots. Of these, the model with the smallest AIC was selected as the optimal model. This approach selected the correct number of knots in the vast majority of samples, and the mean number of knots selected was close to the true value. As has been observed in other contexts, the BIC was biased towards a low number of knots (oversmoothing), and the GCV criterion tended to undersmooth by choosing a higher number of knots more often (data not shown).

The computational effort required by full cross-validation renders it infeasible in this situation. Nevertheless, the focus of this paper is on providing a method of estimation for a single model, which may be applied within any scheme for determining the optimal number of knots.

4.4 Method

4.4.1 CEM algorithm

Each of the methods that we use to fit the models in the subsequent sections is an application of a CEM algorithm (Marschner, 2014). A CEM algorithm is a general approach in which we consider a finite family of complete-data models, indexed by $t \in \mathcal{T}$, each of which has a parameter space $\Theta(t)$ that is a subset of the parameter space Θ for the model of interest, such that

$$\bigcup_{t \in \mathcal{T}} \Theta(t) = \Theta. \tag{4.8}$$

The complete-data models are defined such that an Expectation–Maximisation (EM) algorithm (Dempster, Laird, and Rubin, 1977) can be used to find the constrained maximum of the likelihood $\hat{\boldsymbol{\theta}}(t)$, within each $\Theta(t)$. Then, due to (4.8), the $\hat{\boldsymbol{\theta}}(t)$ that attains the highest likelihood is the MLE $\hat{\boldsymbol{\theta}}$.

When the model of interest is a GLM, the complete-data models typically impose some constraint on the individual model parameters, such as non-negativity. In the case of GAMs as described below, we use augmented complete-data models, where the effect of such constraints on the spline coefficients is to constrain the shape of the smooth curves f_c. The overall MLE will then correspond to one of these constrained maxima. We now describe the details for each of the specific models of interest.

4.4.2 Adjusted rate differences

Adjusted rate differences can be estimated by fitting an identity-link Poisson GAM. Specifically, if N_i is the period of time over which Y_i events were observed, and we assume $Y_i \sim \text{Poisson}(N_i\lambda_i)$, then λ_i is the event rate for an individual with covariate vector $(\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i)$.

With link function $g(\lambda) = \lambda$, the absolute rate difference is simply the difference between two linear functions. Keeping all other covariates constant, the adjusted rate difference associated with changing the a^{th} categorical covariate from level u_{ia} to u_{ja} is $\alpha_a(u_{ja}) - \alpha_a(u_{ia})$. Likewise, the adjusted rate difference for a one-unit increase in the b^{th} linear covariate is β_b.

The Poisson means must be non-negative, and so the parameter space within which the MLE $\hat{\boldsymbol{\theta}}$ must lie is

$$\Theta = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}; \boldsymbol{\theta}) \geq 0, (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}\},$$

where $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{W}$ denote the covariate spaces defined by the Cartesian products of the observed ranges of covariate values:

$$\mathcal{U} = \prod_{a=1}^{A} \{1, \ldots, k_a\}, \quad \mathcal{V} = \prod_{b=1}^{B} [v_b^{(0)}, v_b^{(1)}], \quad \mathcal{W} = \prod_{c=1}^{C} [w_c^{(0)}, w_c^{(1)}]. \tag{4.9}$$

For $C = 0$, Marschner (2010) has described a stable CEM algorithm to find the MLE $\hat{\boldsymbol{\theta}} \in \Theta$. This is achieved by partitioning the parameter space into a sequence of distinct constrained parameter spaces $\Theta'(\boldsymbol{r}, \boldsymbol{s})$, in which $(\boldsymbol{r}, \boldsymbol{s}) = (r_1, \ldots, r_A, s_1, \ldots, s_B)$ is the covariate vector associated with the minimum fitted Poisson mean.

For a particular choice of $\boldsymbol{r}$ and $\boldsymbol{s}$, a complete-data model is defined, consisting of

$A + B + 1$ independent latent random variables underlying each observed random variable Y_i. These latent variables have Poisson distributions with means that each depend on just one of the model parameters. Using this separation of parameters, an EM algorithm is then defined to find the constrained MLE $\hat{\boldsymbol{\theta}}(\boldsymbol{r}, \boldsymbol{s}) \in \Theta'(\boldsymbol{r}, \boldsymbol{s})$.

By considering all possible reference vectors $(\boldsymbol{r}, \boldsymbol{s})$ that could result in the minimum fitted mean, and finding the MLE within each corresponding constrained parameter space, we can be sure that one of these constrained MLEs will be the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$.

For $C > 0$, it is straightforward to extend this idea to handle smooth semi-parametric components. The reference vector is expanded to include $\boldsymbol{t} = (t_1, \ldots, t_C)$, where each $t_c \in \{1, \ldots, D_c\}$. For a particular choice of $\boldsymbol{t}$, the identifiability constraint $\gamma_{ct_c} = 0$ is applied for each $c = 1, \ldots, C$.

The complete-data model is augmented with $\sum_c (D_c - 1)$ latent Poisson random variables $Y_i^{(cd)}$, each having an expected value that depends on one of the remaining parameters, that is,

$$Y_i^{(cd)} \sim \text{Poisson}(N_i \gamma_{cd} B_{cd}(w_{ic})), \qquad d \neq t_c. \tag{4.10}$$

If the basis function values $B_{cd}(w_{ic})$ are viewed as additional continuous covariates, this augmented complete-data model has the same form as that used by Marschner (2010). Thus the EM algorithm can be applied directly to find its MLE $\hat{\boldsymbol{\theta}}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t})$, noting that due to (4.10), each γ_{cd} is constrained to be non-negative.

As discussed in Section 4.3.2, the non-negativity constraints on the γ_{cd} force each estimated f_c to belong to $\mathcal{B}_c^+(t_c)$ as defined in (4.5). The constrained MLE will then belong to the parameter space

$$\Theta(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t}) = \Theta(\boldsymbol{r}, \boldsymbol{s}) \bigcap \left\{ \bigcap_{c=1}^{C} \{\boldsymbol{\theta} : f_c \in \mathcal{B}_c^+(t_c)\} \right\},$$

where

$$\Theta(\boldsymbol{r}, \boldsymbol{s}) = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \geq \Lambda(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{w}) \geq 0, (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}\}.$$

Let $\mathcal{R}$ denote the set of all possible choices for $\boldsymbol{r}$, $\mathcal{S}$ the set of possible choices for $\boldsymbol{s}$

and $\mathcal{T}$ the set of possible choices for $\boldsymbol{t}$. Then due to (4.6), we have that

$$\Theta = \bigcup_{\boldsymbol{r}\in\mathcal{R}} \bigcup_{\boldsymbol{s}\in\mathcal{S}} \bigcup_{\boldsymbol{t}\in\mathcal{T}} \Theta(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t}).$$

We have thus defined a CEM algorithm for finding the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$, which will be the constrained estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t})$ associated with the highest likelihood.

This approach can be implemented directly using the existing CEM algorithm for identity-link Poisson GLMs. Using this algorithm, it is straightforward to apply non-negativity constraints to the coefficients of some continuous covariates by considering only one of the two possible reference levels for that covariate.

For a particular choice of $\boldsymbol{t}$, if we apply the CEM algorithm to the categorical and linear covariates as usual, and include the basis function values $B_{cd}(w_{ic})$ (for $d \neq t_c$) as linear covariates with non-negativity constraints, we will find a constrained MLE $\hat{\boldsymbol{\theta}}(\boldsymbol{t}) \in \Theta(\boldsymbol{t})$, where

$$\Theta(\boldsymbol{t}) = \Theta \bigcap \left\{ \bigcap_{c=1}^{C} \{\boldsymbol{\theta} : f_c \in \mathcal{B}_c^+(t_c)\} \right\}.$$

Repeating this for all $\boldsymbol{t} \in \mathcal{T}$ will result in a collection of constrained MLEs $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$, one of which will be the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$. There are a total of $\prod_{c=1}^{C} D_c$ elements in $\mathcal{T}$, so by applying the CEM algorithm for each, the maximum total number of applications of the EM algorithm required to find the MLE is

$$\prod_{a=1}^{A} k_a \times 2^B \times \prod_{c=1}^{C} D_c.$$

The process can be halted if one such application converges to a point in the interior of the constrained parameter space, as we can be sure that this is the overall MLE.

The process of cycling through the parameter space partition is illustrated in Figure 4.3 for a Poisson model with $C = 1$ covariate. Here we have simulated 500 observations with covariate values w_i and means $\lambda_i = 50(\alpha_0 + f(w_i))$, based on a B-spline with two internal knots

$$f(w) = \sum_{d=1}^{5} \gamma_d B_d(w),$$

where the usual dependence on c has been suppressed since $C = 1$. By virtue of the constraint (4.3), one of the γ_d parameters must be set to zero while all others must be non-negative. The specific B-spline used to simulate the data for Figure 4.3 is

$$\alpha_0 + f(w) = 0.1 + 0.2B_1(w_i) + 0.3B_2(w_i) + 0.4B_4(w_i) + 0.3B_5(w_i),$$

which has $\gamma_3 = 0$ and is represented by the red dotted line in each plot. In this example, cycling through the parameter space partition involves computing the constrained MLE for each of the five models corresponding to the constraint $\gamma_d = 0$, $d = 1, \ldots, 5$. Thus, we would expect the constrained MLE with $\gamma_3 = 0$ to yield the overall MLE. This is shown in Figure 4.3(A), while the constrained MLEs obtained by using the other constraints are shown in Figures 4.3(B)–(E). As expected, the log-likelihood for the curve with $\gamma_3 = 0$ is higher than that for the other constrained MLEs (-1378 compared to -1437 through -1548), so it is the overall MLE.

One point to note is that the non-negativity of the coefficients γ_{cd} is a sufficient, but not necessary, condition for the non-negativity of the estimated f_c. Thus, the parameter space over which we search for the MLE is not guaranteed to include all $f_c \in \mathcal{B}_c^+$. However, in practice it would be highly unlikely that the true relationship we are modelling is exactly a quadratic spline on the selected knot sequence $\boldsymbol{\tau}_c$, and so if we find a MLE that is restricted by this deficiency, we can modify $\boldsymbol{\tau}_c$ in order to achieve a better fit. In testing, we found that simply including an additional knot close to the minimum of the curve will generally resolve this issue.

4.4.3 Adjusted relative risks

With binomial data we have $Y_i \sim \text{Bin}(N_i, \lambda_i)$, where N_i is the number of independent trials over which Y_i binary events were observed, and λ_i is the constant event probability, or risk, at each trial. The adjusted relative risk is the ratio of event probabilities associated with a change in one covariate, keeping the others constant.

With $g(\lambda) = \log(\lambda)$, the risk is $\lambda_i = \exp\{\Lambda(\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{w}_i; \boldsymbol{\theta})\}$, and so the relative risk is a ratio of two exponential functions. Keeping the other covariates constant, the adjusted relative risk associated with a change in the a^{th} categorical covariate from u_{ia} to u_{ja} is then $\exp\{\alpha_a(u_{ja}) - \alpha_a(u_{ia})\}$. Similarly, the adjusted relative risk for a one-unit increase

in the b^{th} linear covariate is $\exp(\beta_b)$.

The fitted risks must lie within $[0, 1]$, and so the fitted linear predictors must be strictly non-positive. That is, the parameter space for this model is

$$\Theta = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}; \boldsymbol{\theta}) \leq 0, (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}\},$$

where $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{W}$ are as defined in (4.9).

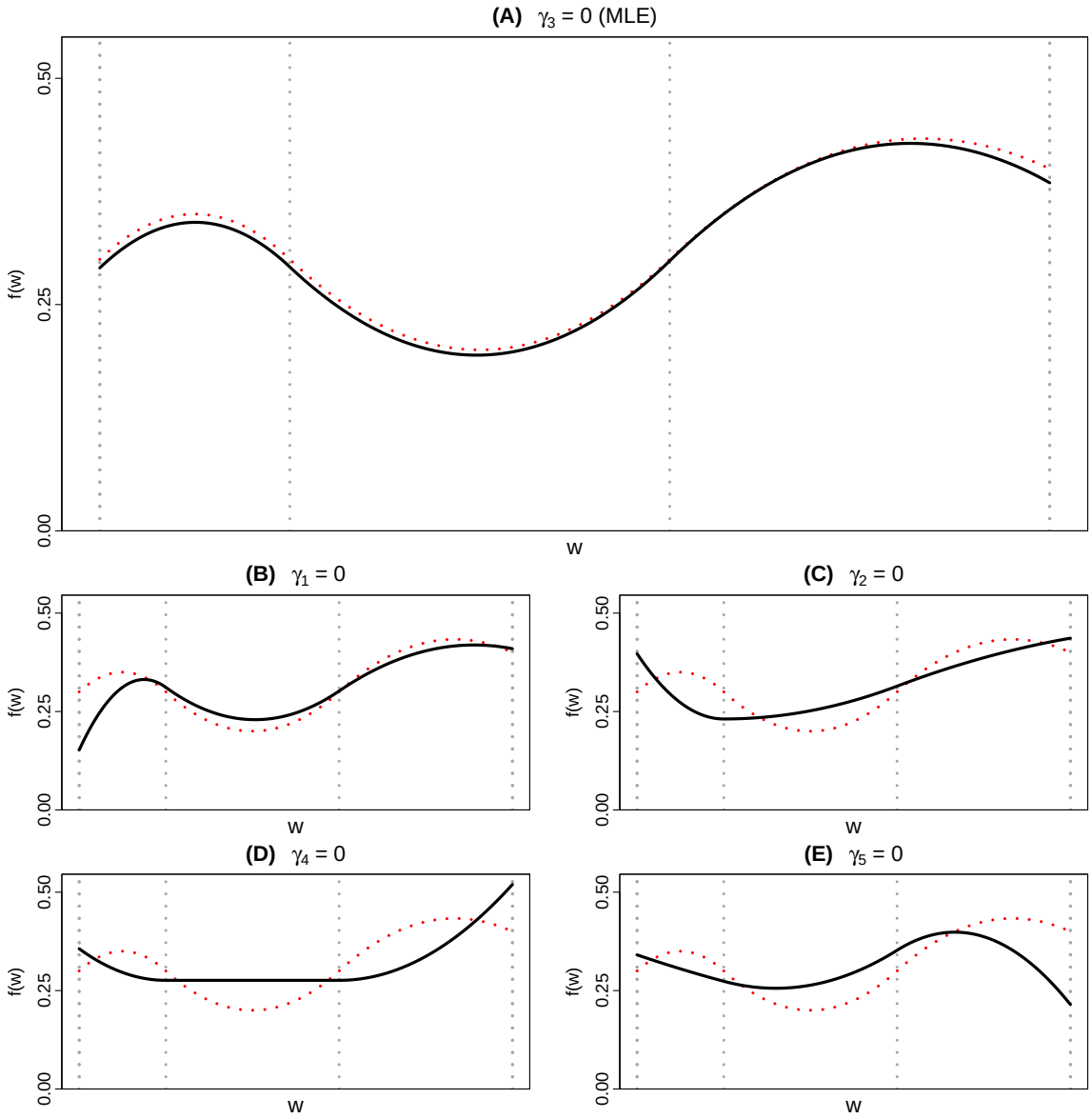


Figure 4.3: Illustration of the effect of parameter constraints on the smooth curve in a simulated Poisson data set. The red dotted line represents the true underlying rate and the grey dotted lines denote the knot locations. The black solid line in each panel is the MLE under the identifiability constraint shown in the heading, with all other parameters constrained to be non-negative. The MLE is obtained with the constraint $\gamma_3 = 0$.

With $C = 0$, Marschner and Gillett (2012) provide a CEM algorithm for finding the MLE $\hat{\boldsymbol{\theta}} \in \Theta$. The approach is similar to that used for the identity-link Poisson GLM: for a given reference vector $(\boldsymbol{r}, \boldsymbol{s})$, the model is reparameterised and a complete-data model is defined, which imposes non-positivity constraints on each of the transformed parameters. An EM algorithm is used to find the MLE $\hat{\boldsymbol{\theta}}(\boldsymbol{r}, \boldsymbol{s})$ of the complete-data model, which is constrained to lie in the parameter space

$$\Theta'(\boldsymbol{r}, \boldsymbol{s}) = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\theta}) \leq \Lambda(\boldsymbol{r}, \boldsymbol{s}; \boldsymbol{\theta}) \leq 0, (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}\}.$$

That is, $(\boldsymbol{r}, \boldsymbol{s})$ is the covariate vector associated with the maximum linear predictor, which is constrained to be non-positive.

The union of these constrained parameter spaces over all possible choices of $(\boldsymbol{r}, \boldsymbol{s})$ is Θ, and so the constrained MLE with the highest likelihood will be the overall MLE. It is straightforward to apply non-positivity constraints to individual parameters.

Extension to include smooth terms follows an analogous approach to that used in Section 4.4.2. For a particular $\boldsymbol{t} = (t_1, \ldots, t_C)$, we set $\gamma_{ct_c} = 0$ and apply the CEM algorithm, including the remaining basis function values $B_{cd}(w_{ic})$ as continuous covariates with non-positivity constraints. Then each $f_c \in \mathcal{B}_c^-(t_c)$, the space of non-positive B-splines that have their shape constrained by the choice of γ_{ct_c} as the largest coefficient. The resulting estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ is the constrained MLE over the parameter space

$$\Theta(\boldsymbol{t}) = \Theta \bigcap \left\{ \bigcap_{c=1}^{C} \{\boldsymbol{\theta} : f_c \in \mathcal{B}_c^-(t_c)\} \right\}.$$

By considering each possible $\boldsymbol{t} \in \mathcal{T}$, and hence all possible locations for the maximum of each f_c, we find a collection of constrained MLEs $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$, and the one with the highest likelihood is then the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$.

Again, the sequence may stop early if a maximum in the interior of the parameter space is identified. Furthermore, the same caution also applies here as it did for the identity-link Poisson model: non-positivity of the coefficients is only a sufficient condition for non-positivity of the linear predictor, so we are not searching the entire parameter space where $f_c \in \mathcal{B}_c^-$. As before, this may be remedied by adjusting the choice of $\boldsymbol{\tau}_c$ if necessary.

4.4.4 Adjusted risk differences

Adjusted risk differences can be estimated using a binomial GAM with identity link, that is, $Y_i \sim \text{Bin}(N_i, \lambda_i)$ and $g(\lambda) = \lambda$. The adjusted risk difference associated with a change in the a^{th} categorical covariate from u_{ia} to u_{ja} is $\alpha_a(u_{ja}) - \alpha_a(u_{ia})$. Similarly, β_b represents the adjusted risk difference for a one-unit increase in the b^{th} linear covariate.

Again we require that the probabilities λ lie within $[0, 1]$, but now the parameter space Θ simultaneously imposes both lower and upper boundaries on the linear predictors, that is,

$$\Theta = \{\boldsymbol{\theta} : 0 \leq \Lambda(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}; \boldsymbol{\theta}) \leq 1, (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}\}.$$

With $C = 0$, Donoghoe and Marschner (2014) have presented an approach for finding the MLE $\hat{\boldsymbol{\theta}} \in \Theta$. It exploits the multinomial–Poisson transformation (Baker, 1994) to convert the model into an equivalent identity-link Poisson fit, and the CEM algorithm of Marschner (2010) can then be applied to the transformed data to find the MLE. As with the other CEM algorithms, non-negativity constraints can be applied to some of the parameters, which is achieved by imposing such constraints when employing the identity-link Poisson CEM algorithm.

However, the inclusion of the semi-parametric terms cannot proceed in exactly the same way as in Sections 4.4.2 and 4.4.3. The covariate space considered for continuous covariates entered into the identity-link binomial CEM algorithm is the Cartesian product of the observed ranges of those covariates, that is, $\mathcal{V}$ in (4.9). So if we include the observed B-spline values $B_{cd}(w_{ic})$ as continuous covariates, the covariate space will include vectors in which more than three of the basis functions for a given c are non-zero. By definition, this cannot correspond to the B-spline values for any w. Since the parameter space constraints are imposed for all points in the covariate space, using a larger covariate space than is necessary means that the parameter space is overly restrictive, and may not include the overall MLE.

Instead we must use a slightly different approach, based on the ordering of the B-spline coefficients. We begin by choosing $\boldsymbol{t} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_C)$, where each $\boldsymbol{t}_c = (t_{c1}, \ldots, t_{cD_c})$ is now a vector containing some permutation of $\{1, \ldots, D_c\}$.

Recall that if we define new basis functions B_{cd}^{m+} as described in Section 4.3.3, non-negativity constraints on the coefficients δ_{cd} will impose monotonicity on the coefficients

of the original basis functions, that is, $\gamma_{c1} \leq \cdots \leq \gamma_{cD_c}$. Similarly, for a particular permutation $\boldsymbol{t}_c$, we can define

$$B_{cd}^{(\boldsymbol{t}_c)}(w) = \sum_{e=d}^{D_c} B_{ct_{ce}}(w), \qquad d = 2, \ldots, D_c,$$

and non-negativity constraints on the associated coefficients will impose an order restriction on the original coefficients, that is, $0 = \gamma_{ct_{c1}} \leq \cdots \leq \gamma_{ct_{cD_c}}$.

We denote by $\mathcal{B}_c^{(\boldsymbol{t}_c)+}$ the subspace of $\mathcal{B}_c^+$ in which the coefficients are ordered in this way, and note that the overall MLE must correspond to one such permutation. If for a particular $\boldsymbol{t}$ we enter the new basis function values $B_{cd}^{(\boldsymbol{t}_c)}(w_{ic})$ as continuous covariates into the identity-link binomial CEM algorithm, and impose non-negativity constraints on their coefficients, the resulting estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ will be the MLE for the parameter space

$$\Theta(\boldsymbol{t}) = \Theta \bigcap \left\{ \bigcap_{c=1}^{C} \{\boldsymbol{\theta} : f_c \in \mathcal{B}_c^{(\boldsymbol{t}_c)+}\} \right\}.$$

There are $D_c!$ possible choices for each $\boldsymbol{t}_c$, and hence $\prod_{c=1}^{C} D_c!$ possible choices for the vector $\boldsymbol{t}$. In order to find the overall MLE $\hat{\boldsymbol{\theta}} \in \Theta$, we find the constrained MLE $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ for each possible choice of $\boldsymbol{t}$, and choose the one with the highest likelihood.

As with the other models, we may stop early if we find a constrained MLE in the interior of the parameter space, and there are various strategies to identify the parameterisations that are more likely to contain this MLE in order to potentially reduce computing time (Marschner, 2014).

4.4.5 Monotonic smooth regression

Imposing a monotonicity restriction on one or more of the smooth curves is straightforward for all of the models examined so far. As discussed in Section 4.3.3, a sufficient condition for f_c to be monotonically non-decreasing is that the B-spline coefficients themselves are monotonically non-decreasing, that is, $\gamma_{c1} \leq \cdots \leq \gamma_{cD_c}$. Because only one possible choice for the value t_c or the vector $\boldsymbol{t}_c$ needs to be considered, adding a smooth monotonic covariate to a model requires no additional applications of the EM algorithm.

For the identity-link binomial model in Section 4.4.4, applying this constraint is trivial.

In iterating through the possible choices of $\boldsymbol{t}$, we need only to consider $\boldsymbol{t}_c = (1, \ldots, D_c)$ rather than all possible permutations, giving $0 = \gamma_{c1} \leq \cdots \leq \gamma_{cD_c}$.

For the other models, we employ the methods discussed in Section 4.3.3. In an identity-link Poisson model, we replace the B-spline function values $B_{cd}(w_{ic})$ by their monotonic B-spline equivalents, $B_{cd}^{m+}(w_{ic})$, and enter these into the CEM algorithm as continuous covariates with non-negativity constraints on the associated coefficients δ_{cd}. The resulting estimate will have $0 = \gamma_{c1} \leq \cdots \leq \gamma_{cD_c}$, as desired.

For the log-link binomial model we proceed similarly, replacing $B_{cd}(w_{ic})$ by $B_{cd}^{m-}(w_{ic})$ as defined in (4.7). These are entered into the CEM algorithm as continuous covariates with non-positivity constraints, ensuring $\gamma_{c1} \leq \cdots \leq \gamma_{cD_c} = 0$.

It is important to note that the monotonicity of the coefficients is only a sufficient condition for the monotonicity of the resulting smooth function. Hence the function space over which we search for the MLE $\hat{f}_c$ is only a subset of the space of monotonic functions on a B-spline basis. However, when Tutz and Leitenstorfer (2007) used boosting techniques to impose constraints on the monotonicity of the function rather than only the coefficients, they found "much higher computational costs without much effect on performance".

4.5 Application

The ASSENT-2 study (ASSENT-2 Investigators, 1999) was a randomised clinical trial designed to assess the safety and efficacy of tenecteplase versus alteplase in 16,949 patients treated within 6 hours of an acute myocardial infarction (MI). The primary outcome was 30-day mortality after randomisation, and the primary analysis of this outcome showed that the two treatments were equivalent.

To demonstrate our method, we undertake a risk factor analysis, in which the risk of death is modelled in terms of a semi-parametric age effect and three categorical covariates: MI severity (Killip class I, II or III/IV), treatment delay (< 2, 2–4, > 4 hours) and geographic region (Western countries, Latin America or Eastern Europe). In view of the natural relationship between age and death, the age-specific risk is constrained to be monotonically non-decreasing. Figure 4.4 shows the AIC for models with different numbers of knots, using both log and identity link functions.

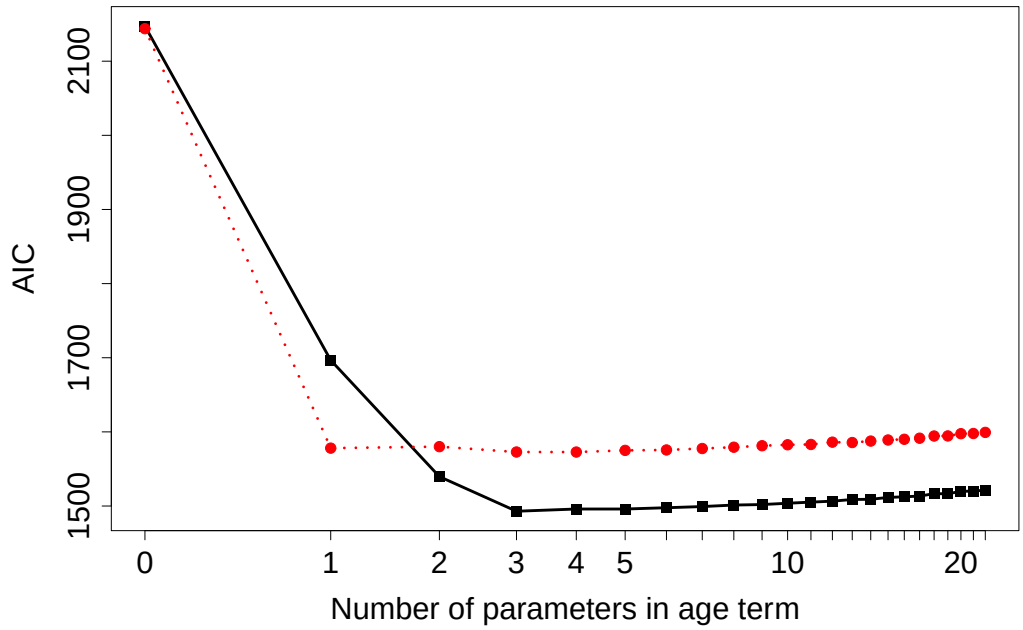


FIGURE 4.4: *AIC* of the fitted binomial model in the ASSENT-2 study with identity (black, solid line) and log (red, dotted line) links, for a varying number of parameters associated with the age covariate.

Figure 4.4 shows that the inclusion of a linear age term (1 parameter) improves the *AIC* considerably for both the log-link and identity-link models. In the risk difference model, allowing the age term to be quadratic (2 parameters) and semi-parametric with one internal knot (3 parameters) further reduces the *AIC* substantially, but beyond this, the penalty of additional parameters overrides the small improvements in fit. In the relative risk model, additional flexibility in the age term does not give vastly superior *AIC* values when compared to the linear age model, although the best in terms of *AIC* is a model with two internal knots, and hence 4 parameters for the semi-parametric age term. The fitted risks by age for all 27 groups ($3 \times 3 \times 3$ levels of the categorical covariates) for the best model with each link are shown in Figure 4.5. These results show that the identity-link model provides a better fit to the data, suggesting that risk differences are a more appropriate measure than relative risks for this data set.

Although a monotonic dependence on age is natural for mortality, in this particular analysis it has little effect on the fitted model. Figure 4.6 displays the fitted age-specific regression functions for both models, showing that the fitted function is virtually identical for the identity-link model with and without monotonicity, and is identical for the

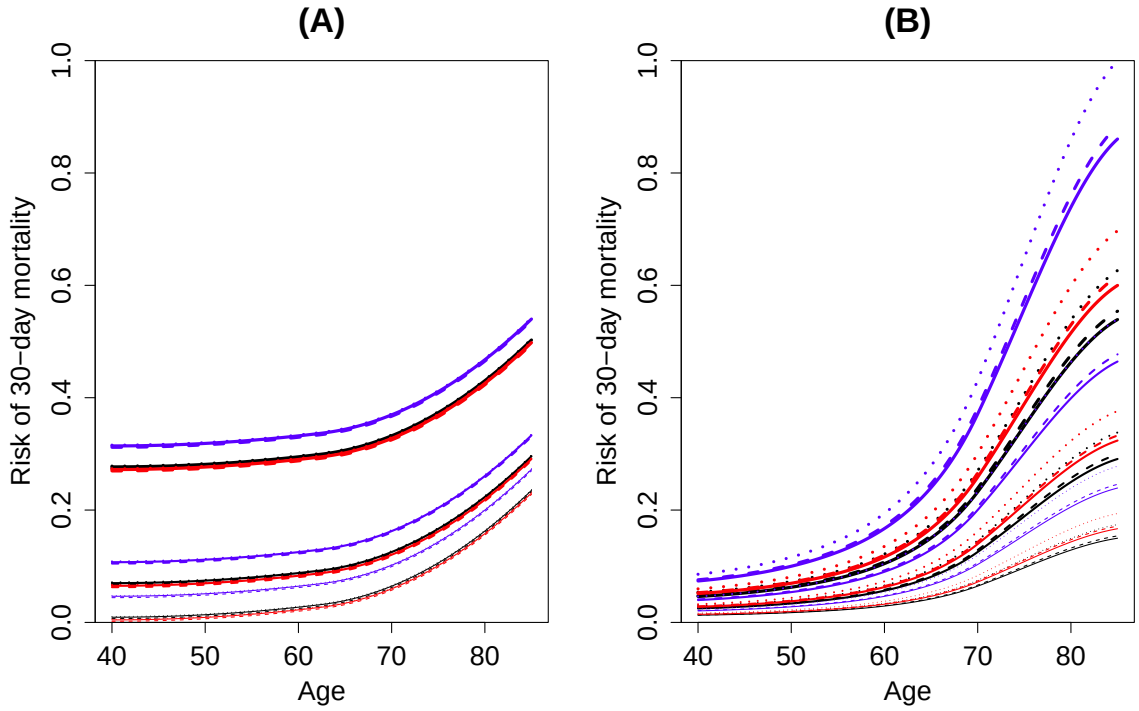


Figure 4.5: Fitted risk of 30-day mortality by age in the ASSENT-2 study for different combinations of MI severity (thin: I, medium: II, thick: III/IV), treatment delay (solid: < 2 hours, dashed: 2–4 hours, dotted: > 4 hours) and region (black: Western, red: Latin America, blue: Eastern Europe), for (A) identity-link and (B) log-link models.

log-link model.

Also shown in Figure 4.6 is a comparison of two approaches for confidence interval estimation. The first approach, shown by the shaded regions, uses asymptotic normality and the information matrix evaluated at the MLE. The information matrix is obtained from the expected second derivative matrix corresponding to the binomial log-likelihood function with probabilities specified by (4.1) and (4.2), with either the identity or log link function. The second approach, shown by the dashed lines, uses bootstrapping with pointwise confidence intervals determined using the percentile method. We used 1000 bootstrap resamples with replacement, and due to the stability of our fitting methods, convergence to the MLE was achieved in every resample. It can be seen that the two methods of confidence interval estimation show very close agreement, which provides some level of support for their use in this analysis.

Adjusted rate differences can also be estimated for this data by using our method for an identity-link Poisson model, with semi-parametric adjustment for age. Because in this case all of the patients were observed for the same period of time, the parameter

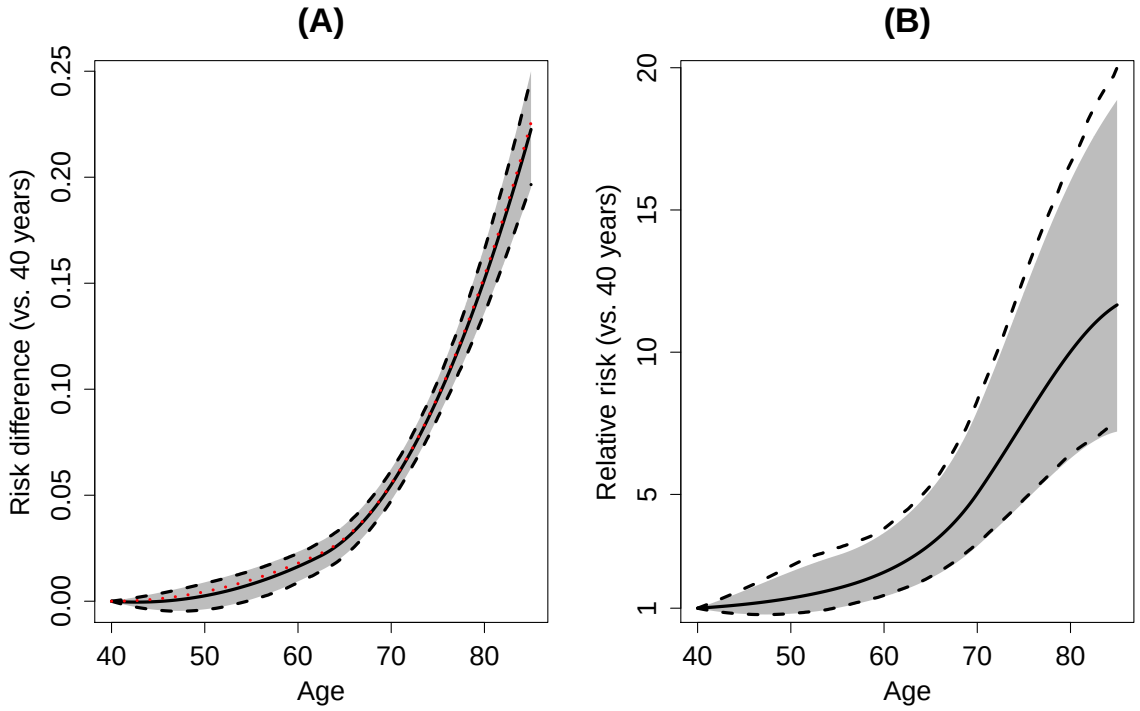


FIGURE 4.6: (A) Adjusted risk difference and (B) adjusted relative risk associated with age (versus 40 years), with pointwise 95% confidence intervals, estimated using the information matrix (shaded) and bootstrap resampling (dashed lines). The dotted red line in panel (A) shows the estimated risk difference when a monotonicity constraint is applied.

TABLE 4.1: Parameter estimates from identity-link binomial (risk difference) and identity-link Poisson (rate difference) models on the ASSENT-2 data.

Covariate	Risk difference	Rate difference
Severity (vs. I)		
II	0.0607	0.0605
III / IV	0.2676	0.2692
Treatment delay (vs. < 2 hours)		
2–4 hours	-0.0024	-0.0025
> 4 hours	0.0014	0.0010
Region (vs. Western)		
Latin America	-0.0047	-0.0050
Eastern Europe	0.0373	0.0359
Age (vs. 40 years)		
50 years	0.0045	0.0045
60 years	0.0179	0.0178
70 years	0.0553	0.0553
80 years	0.1534	0.1541

estimates are very similar to those from the identity-link binomial model; however, they have a different interpretation: they are the absolute change in the rate of death per patient-month. A comparison of the parameter estimates from the two models is

shown in Table 4.1.

4.6 Other methods

The methods described in this paper are fully implemented in R packages `addreg` (Donoghoe, 2015a) and `logbin` (Donoghoe, 2015b), available from the Comprehensive R Archive Network (CRAN). There are three other notable R packages that provide methods for fitting GAMs: `gam`, `gamlss` and `mgcv`. Each of these allow the models with non-canonical link functions discussed in this paper; however, all employ iterative algorithms involving variants of Fisher scoring or Newton–Raphson, making them subject to instability.

The `gam` function in the `gam` package in R (Hastie, 2013) fits GAMs using cubic smoothing splines by employing a local scoring algorithm. This consists of a backfitting (Gauss–Seidel) algorithm for fitting the non-parametric parts of the model within a Newton–Raphson step for updating the parametric parts (Hastie, 1992). The inner loop can be shown to always converge, but the outer loop is only guaranteed to do so if some form of step-size optimisation is performed (Hastie and Tibshirani, 1990, p. 151). However, the implementation in R does not include any option for step-size modification, and additionally there is no check for the validity of the fitted means. This means that convergence is not guaranteed, and when the method does converge, it may be to a value outside the parameter space. For the 1000 bootstrap samples used to produce the confidence intervals in Figure 4.6, the algorithm failed to converge to a valid solution in nearly half (49.1%) of the samples for the identity-link model. For the log-link model, the algorithm converged in all 1000 samples, but some of the fitted risks exceeded 1 in every case.

The `gamlss` package (Rigby and Stasinopoulos, 2005) provides a method for non-parametric modelling of various parameters of the distribution, including the mean. Similarly to `gam`, its fitting algorithm uses backfitting iterations for the non-parametric parts within Newton–Raphson steps that update the parameter estimates. Unlike `gam`, the user is able to specify the step length for updating parameter estimates, but the function terminates with an error if the update produces invalid fitted values. For the relative risk model fitted on the bootstrap samples from Section 4.5, `gamlss` converged

in only 57 samples when the default step size was used. Convergence in all 1000 samples was achieved when the step size was made sufficiently small. For the risk difference model however, `gamlss` did not converge in any of the bootstrap samples regardless of the step size chosen, due to the error caused by violation of the parameter space constraint.

The `mgcv` package (Wood, 2011) provides a flexible `gam` function, allowing for a wide variety of smoothers, as well as methods for automatic selection of the level of smoothing. The fitting method uses a penalised iteratively reweighted least squares algorithm (Wood, 2011), which is not guaranteed to converge, but step-halving is invoked if the penalised deviance increases markedly between iterations, or the estimates move outside the parameter space. Additionally, the user can specify a ridge regression penalty to assist with convergence issues caused by unidentifiable estimates. In order to directly compare the performance of `mgcv`'s `gam` function to our method, we fitted identical unpenalised B-spline models to the ASSENT-2 bootstrap data. We found that `gam` achieved stable convergence whenever the MLE was in the interior of the parameter space, but convergence problems were possible when the MLE was on the parameter space boundary, particularly for the identity-link model. The nature of these convergence problems was dependent on the version of `mgcv` that was used. In particular, when using version 1.7 we found that convergence could occur to a sub-optimal boundary point, while in version 1.8 we found that the algorithm could fail to declare convergence when the estimates reached the MLE. This behaviour occurred in a large proportion of our bootstrap replications and persisted even when a stricter convergence criterion and a greater number of iterations were used.

None of these methods support monotonicity constraints on the smooth curves. The `GMBBoost` (Leitenstorfer and Tutz, 2007) and `GMonBoost` (Tutz and Leitenstorfer, 2007) methods employ likelihood-based boosting techniques to fit GAMs with monotonicity constraints, but in the current implementation only canonical-link models are allowed, so we cannot compare them with our approach.

Overall, this discussion illustrates that although there are other approaches that could potentially be used for semi-parametric modelling of rate differences, risk differences and relative risks, numerical instability is often an issue. Furthermore, monotonicity constraints for non-canonical models are not available in existing software. The stability

and flexibility of our method therefore means that it is a useful addition to existing GAM methodology.

4.7 Discussion

We have presented a method for smooth semi-parametric adjustment of rate differences, relative risks and risk differences. In general, this can be achieved by using GAMs; however, these effect measures require non-standard link functions and the usual fitting algorithms can fail to converge to the MLE. Our method avoids this by employing variants of existing stable CEM algorithms for fully parametric versions of these models, using B-spline basis functions for the smooth components.

The method is itself a CEM algorithm, and relies on the fact that the EM algorithm will always converge to the MLE within a constrained parameter space. Each constrained parameter space is defined by placing a restriction on the shape of the smooth curve, and by applying the algorithm for each constrained parameter space we are guaranteed to find the overall MLE.

We applied our method to data from the ASSENT-2 clinical trial, showing that semi-parametric adjustment for age provided a better fit than entering age as a linear term in both risk difference and relative risk models. Furthermore, the stability of our fitting algorithm allowed us to use bootstrap resampling to estimate confidence intervals for the semi-parametric relationship. Adjusted rate differences can also be estimated using an identity-link Poisson model.

The calculations required at each iteration of the EM algorithms for each method presented here are very simple, although the EM algorithm may take a large number of iterations to converge. The overall computational time required by these methods depends on the number of parameters that must be estimated in a particular model. The models presented in Figure 4.5 required approximately 3 and 2 minutes, respectively, to find the MLE on a 3.4 GHz processor. One potential method for reducing the computational time is to exploit the fact that the EM algorithms for each parameterisation are independent, and could be conducted in parallel on a multi-core processor. Other techniques for speeding up convergence of CEM algorithms have been discussed by Marschner (2014).

Further adaptations of our approach are possible. For example, Marschner, Gillett, and O'Connell (2012) have presented an extension of the CEM algorithm for the identity-link Poisson model, in which additional categorical stratification factors have a multiplicative effect on the Poisson means. The algorithm is similar to that for the additive model, but each constrained MLE is found by using an Expectation–Conditional Maximisation (ECM) algorithm. It is straightforward to impose non-negativity constraints on the additive parameters, so we can extend this method to allow smooth semi-parametric components by using the same approach described in Section 4.4.2. An alternative approach to incorporating both additive and multiplicative effects in the same model was provided by the LEXPIT model of Kovalchik et al. (2013). We anticipate that our approach may be useful in extending this model semi-parametrically, although we have not yet investigated this.

Most applications of GAMs use penalised likelihood to allow for flexibility while lessening the tendency to overfit. However, any reasonable penalty term will cause the M-step of the EM algorithm to lose its parameter separation and become a multi-dimensional maximisation problem. Marschner and Gillett (2012) proposed a solution to this for log-binomial models by employing the one-step-late algorithm of Green (1990). Here, the M-step is modified such that parameters associated with the penalty term are replaced by their current estimates. However, this is no longer an EM algorithm, and does not guarantee that the parameter estimates will remain in the parameter space, or even that the likelihood will increase at each step. This can be remedied by a process similar to step-halving, whereby if the new estimate has lower likelihood or is outside the parameter space, we replace it with an estimate between the unpenalised and penalised updates, moving closer to the unpenalised estimate until the conditions are met. Nonetheless, while a penalised likelihood version of our method would be possible, the simple spline-based model used here is likely to provide sufficient flexibility in practice.

Rate differences, relative risks and risk differences are useful in biostatistical settings to provide effect size measures in randomised trials and epidemiological studies. However, the GLMs and GAMs used to estimate these effects are also used in other areas. For example, the identity-link Poisson model has been recently used in an ecological study (Stjernman et al., 2013), the log-link binomial model in a study of socioeconomic status

(Hiyoshi et al., 2013), and the identity-link binomial model in a study of international politics (Berger et al., 2013). This suggests that our method may also have wide applicability outside biostatistics.

Acknowledgements

The authors thank the ASSENT-2 investigators for providing data to the National Health and Medical Research Council Clinical Trials Centre, and the anonymous reviewers for their constructive comments to help improve this paper.

Funding

This research was supported by the Australian Research Council (DP110101254).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716–723. DOI: 10.1109/tac.1974.1100705.

ASSENT-2 Investigators (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *The Lancet* **354**(9180): 716–722. DOI: 10.1016/S0140-6736(99)07403-6.

Baker, S. G. (1994). The multinomial–Poisson transformation. *The Statistician* **43**(4): 495–504. DOI: 10.2307/2348134.

Berger, D., A. Corvalan, W. Easterly, and S. Satyanath (2013). Do superpower interventions have short and long term consequences for democracy? *Journal of Comparative Economics* **41**(1): 22–34. DOI: 10.1016/j.jce.2013.01.004.

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer-Verlag.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1): 1–38.

Donoghoe, M. W. (2015a). *addreg: Additive Regression for Discrete Data*. R package version 2.0. URL: http://CRAN.R-project.org/package=addreg.

Donoghoe, M. W. (2015b). *logbin: Relative Risk Regression Using the Log-Binomial Model*. R package version 1.2. URL: http://CRAN.R-project.org/package=logbin.

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* **80**: 184–196. DOI: 10.1016/j.csda.2014.06.019.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**(2): 89–121. DOI: 10.1214/ss/1038425655.

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**(1): 3–39. DOI: 10.2307/1270359.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **52**(3): 443–452.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. 1st ed. London: Chapman & Hall.

Hastie, T. J. (1992). "Generalized additive models". *Statistical Models in S*. ed. by J. M. Chambers and T. J. Hastie. Pacific Grove, CA: Wadsworth & Brooks/Cole. Chap. 7: pp. 249–307.

Hastie, T. J. (2013). *gam: Generalized Additive Models*. R package version 1.09. URL: http://CRAN.R-project.org/package=gam.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. 1st ed. London: Chapman & Hall.

Hiyoshi, A., Y. Fukuda, M. J. Shipley, M. Bartley, and E. J. Brunner (2013). A new theory-based social classification in Japan and its validation using historically collected information. *Social Science & Medicine* **87**: 84–92. DOI: 10.1016/j.socscimed.2013.03.021.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(2): 271–293. DOI: 10.1111/1467-9868.00125.

Kovalchik, S. A., R. Varadhan, B. Fetterman, N. E. Poitras, S. Wacholder, and H. A. Katki (2013). A general binomial regression model to estimate standardized risk differences from binary response data. *Statistics in Medicine* **32**(5): 808–821. DOI: 10.1002/sim.5553.

Leitenstorfer, F. and G. Tutz (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8**(3): 654–673. DOI: 10.1093/biostatistics/kxl036.

Malloy, E. J., D. Spiegelman, and E. A. Eisen (2009). Comparing measures of model selection for penalized splines in Cox models. *Computational Statistics and Data Analysis* **53**(7): 2605–2616. DOI: 10.1016/j.csda.2008.12.008.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**(3): 666–683. DOI: 10.1198/jcgs.2010.09127.

Marschner, I. C. (2011). glm2: Fitting generalized linear models with convergence problems. *The R Journal* **3**(2): 12–15.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* **24**(6): 921–940. DOI: 10.1007/s11222-013-9411-7.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics* **13**(1): 179–192. DOI: 10.1093/biostatistics/kxr030.

Marschner, I. C., A. C. Gillett, and R. L. O'Connell (2012). Stratified additive Poisson models: Computational methods and applications in clinical epidemiology. *Computational Statistics and Data Analysis* **56**(5): 1115–1130. DOI: 10.1016/j.csda.2011.08.002.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**(4): 425–441. DOI: 10.1214/ss/1177012761.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**: 507–554. DOI: 10.1111/j.1467-9876.2005.00510.x.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**(4): 735–757. DOI: 10.1198/106186002853.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. New York: Cambridge University Press: p. 131.

SAS Institute Inc. (2008). *SAS/STAT Software, Version 9.2*. Cary, NC. URL: http://www.sas.com/.

Stjernman, M., M. Green, Å. Lindström, O. Olsson, R. Ottvall, and H. G. Smith (2013). Habitat-specific bird trends and their effect on the Farmland Bird Index. *Ecological Indicators* **24**: 382–391. DOI: 10.1016/j.ecolind.2012.07.016.

Tutz, G. and F. Leitenstorfer (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics* **16**(1): 165–188. DOI: 10.1198/106186007X180949.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(1): 3–36. DOI: 10.1111/j.1467-9868.2010.00749.x.

Yao, F. and T. C. M. Lee (2008). On knot placement for penalized spline regression. *Journal of the Korean Statistical Society* **37**(3): 259–267. DOI: 10.1016/j.jkss.2008.01.003.

5

Additive negative binomial regression

Marschner (2010) has described a stable method for estimating adjusted rate differences by using an additive Poisson regression model, and in Chapter 4, we extended this approach so that flexible smooth regression functions can be included, allowing for semi-parametric adjustment of rate differences.

However, the Poisson model assumes equidispersion, meaning that the conditional variance of the number of events is equal to the conditional mean. In this chapter, we consider the phenomenon that is often observed with real data, where the variance exceeds the mean, known as overdispersion.

We show that a model for the unobserved heterogeneity that leads to overdispersion in count data gives the negative binomial distribution, and so additive negative binomial regression can be used to estimate adjusted rate differences when overdispersion is present. We describe a stable algorithm for maximum likelihood estimation in additive negative binomial models, using repeated applications of an ECME algorithm.

The method also allows for the inclusion of flexible smooth terms, resulting in semi-parametrically adjusted rate differences. We provide an example of its use in estimating the effect of fenofibrate treatment on the rate of laser therapy for retinopathy in data from the FIELD clinical trial.

The content in this chapter makes up a manuscript that has been prepared for submission to a peer reviewed journal:

Donoghoe, M. W. and I. C. Marschner. Additive negative binomial regression for overdispersed count data with application to adjusted rate differences. Manuscript ready for submission.

The manuscript includes supplementary material, which has been reproduced here in Appendices 5.A and 5.B. The R package that implements the method described in this chapter is available online at `http://CRAN.R-project.org/package=addreg`, with its documentation presented in Appendix B of this thesis.

Specific contribution of co-authors: I. C. Marschner assisted with conception of the method, and provided general supervision and feedback on research and writing. The candidate's contribution was at least 90% of the total effort required to produce the article.

Additive negative binomial regression for overdispersed count data with application to adjusted rate differences

Mark W. Donoghoe[1,2], Ian C. Marschner[1,2]

[1] Department of Statistics, Macquarie University, NSW 2109, Australia

[2] NHMRC Clinical Trials Centre, University of Sydney, NSW 2006, Australia

Abstract

Rate differences are an important effect measure in biostatistics and provide an alternative perspective to rate ratios. When the data are event counts observed during an exposure period, adjusted rate differences may be estimated using an identity-link Poisson generalised linear model, also known as additive Poisson regression. A problem with this approach is that the assumption of equality of mean and variance rarely holds in real data, which often show overdispersion. An additive negative binomial model is the natural alternative to account for this, however, standard model-fitting methods are often unable to cope with the constrained parameter space arising from the non-negativity restrictions of the additive model. In this paper, we propose a novel solution to this problem using a variant of the ECME algorithm. Our method provides a reliable way to fit an additive negative binomial regression model and also permits flexible generalisations using semi-parametric regression functions. We illustrate the method using a placebo-controlled clinical trial of fenofibrate treatment in patients with type II diabetes, where the outcome is the number of laser therapy courses administered to treat diabetic retinopathy. An R package is available that implements the proposed method.

Keywords: ECME algorithm · Negative binomial regression · Overdispersion · Rate difference · Semi-parametric regression

5.1 Introduction

In many biostatistical contexts we observe an outcome that counts the number of times an event of interest occurred and often these counts are observed over differing units of exposure, such as differing time periods of follow-up. When a collection of covariates is also available, such data allow us to build a regression model quantifying the effects of the covariates on the rate at which events occur. In this context, the parameters in an additive regression model represent adjusted rate differences, and provide an alternative to the adjusted rate ratios from a multiplicative model.

Rate differences may often be preferable to rate ratios as a measure of effect size in biostatistical applications. In studies evaluating an intervention, a rate difference quantifying the effect of the intervention directly relates to the expected number of events that may be prevented by its use. For example, in vaccine studies the rate difference is called the vaccine-attributable reduction, and can provide better information about the number of infections prevented than the vaccine efficacy, which is a relative effect size (Greenwood, 2005). A similar distinction exists in epidemiological cohort studies where the rate difference is called the attributable risk. Rate differences are also useful in health economics, because they allow quantification of the cost of an intervention per event prevented within a given time period. Furthermore, risk factor models used for prediction and stratification are sometimes best presented in terms of rate differences rather than rate ratios, both from an interpretation perspective and because additive models sometimes fit the data better than multiplicative models.

A common approach for estimating adjusted rate differences is to fit a generalised linear model (GLM) with an identity link function, and assume that the observed counts have a Poisson distribution. We will refer to this analysis as additive Poisson regression. Provided the mean model is correctly specified, the parameter estimates from such a model will be consistent (Winkelmann, 2008, p. 80), but the Poisson constraint that the variance of the count is equal to its expected value often does not hold true in real data. In particular, it is common for count data to exhibit overdispersion, where the variance exceeds the mean (Tang, He, and Tu, 2012).

Overdispersion can result from various causes. Perhaps the most common cause is the presence of unobserved heterogeneity between individuals. Other causes are also

possible, including the so-called positive contagion phenomenon discussed by Bates and Neyman (1952). The presence of overdispersion leads to an excess of both low and high counts compared to those predicted by the Poisson distribution (Cameron and Trivedi, 1998, pp. 98–100). An important consequence of this is that the estimated standard errors from a Poisson model will be underestimates, leading to overly narrow confidence intervals and inflated type I error rates (Gourieroux, Monfort, and Trognon, 1984).

Valid standard errors can be obtained by using a robust sandwich variance estimator (Winkelmann and Zimmermann, 1992), or by using the quasilikelihood approach of Wedderburn (1974), which only requires the mean–variance relationship rather than a full model (Winkelmann, 2008, pp. 91–93). However, because these methods do not employ a distributional model, likelihood-based criteria such as AIC and BIC cannot be applied for model selection (Burnham and Anderson, 2002), and it is not possible to estimate the expected distribution of event counts in a population.

Several alternatives to Poisson GLMs for analysing overdispersed count data have been presented in the literature. The most common approach is to use negative binomial regression, which has been reviewed by various authors; see for example the books by Cameron and Trivedi (1998) and Hilbe (2011). In principle, additive negative binomial regression using an identity link function would therefore provide a method to estimate adjusted rate differences accounting for overdispersion. However, in practice this presents challenges. The model fitting methods used to implement negative binomial regression are generally reliable for multiplicative models, because on the log link scale the parameter space is unconstrained. However, for additive models a restriction must be placed on the possible values of the parameters to ensure that estimated mean counts and rates are non-negative. For this reason, standard gradient-based algorithms for fitting additive negative binomial models often suffer convergence issues and can be unreliable.

The purpose of this paper is to address these challenges by presenting a reliable and flexible approach to additive negative binomial regression and using it to estimate adjusted rate differences in count data subject to overdispersion. Section 5.2 will outline the additive model that we will use based on the negative binomial distribution, and then Section 5.3 will describe how a variant of the Expectation–Conditional

Maximisation–Either (ECME) algorithm can be used to reliably fit this model. We will then extend this in Section 5.4 to allow additional flexibility through smooth semi-parametric regression functions. A detailed illustrative application of the methods will then be provided in Section 5.5, using data from a large clinical trial in type II diabetics where the outcome is the number of laser therapy courses administered to treat diabetic retinopathy.

5.2 Additive negative binomial regression

The negative binomial distribution can be derived in several ways (Johnson, Kotz, and Kemp, 1993), but in this context we motivate its use by introducing multiplicative Gamma-distributed errors into the Poisson model. We begin by assuming that the number of events Y_i for individual i ($i = 1, \ldots, n$) is distributed as

$$Y_i \sim \text{Poisson}(N_i \lambda_i).$$

Here, N_i is the fixed exposure over which Y_i is observed, such that λ_i represents the event rate per unit of exposure for individual i. Typically, N_i will correspond to a fixed time period of observation for individual i, so that λ_i is the event rate per unit of time, although other types of exposure are also possible. When modelling count data rather than rates, we can set $N_i = 1$ for all i so that λ_i is the expected count.

Under an additive regression model, the event rate is a linear function of the covariate vector $\boldsymbol{x}_i$ and its associated parameter vector $\boldsymbol{\theta}^*$, which we denote by

$$\lambda_i = \Lambda(\boldsymbol{x}_i, \boldsymbol{\theta}^*) = \boldsymbol{x}_i \boldsymbol{\theta}^*. \tag{5.1}$$

This is an identity-link Poisson GLM in which component j of $\boldsymbol{\theta}^*$ is the adjusted rate difference per unit change in component j of $\boldsymbol{x}_i$. This model is a natural alternative to the multiplicative Poisson model with log link function, where the (exponentiated) parameters are adjusted rate ratios.

The Poisson model has the restrictive assumption that $\text{Var}(Y_i) = \mathbb{E}(Y_i)$, which is frequently violated in the direction of overdispersion, $\text{Var}(Y_i) > \mathbb{E}(Y_i)$. A natural way to accommodate this overdispersion is to adopt a more general mean–variance

relationship, $\text{Var}(Y_i) = \sigma^2\mathbb{E}(Y_i)$, for $\sigma^2 > 1$. For notational convenience we will rewrite this relationship in the equivalent form $\text{Var}(Y_i) = (1+\phi)\mathbb{E}(Y_i)$, for $\phi > 0$. Then the desired mean–variance relationship can be achieved by generalising the Poisson model such that the mean is perturbed by multiplicative errors,

$$Y_i \mid \eta_i \sim \text{Poisson}(N_i\lambda_i\eta_i),$$

where η_i is a random variable with $\mathbb{E}(\eta_i) = 1$ and $\text{Var}(\eta_i) = \phi/(N_i\lambda_i)$. This can be confirmed by a straightforward application of the law of total variance.

One possible cause of such overdispersion is unobserved heterogeneity in which some individuals are more event-prone than others due to factors that are not captured by the observed covariates. Whatever the cause of the overdispersion, λ_i retains its interpretation as an event rate, and $\boldsymbol{\theta}^*$ its interpretation as a vector of adjusted rate differences, averaged over the population.

Specification of the model is completed by assuming a distribution for η_i. A commonly used approach is to assume that η_i is Gamma-distributed, which dates back to Greenwood and Yule (1920). This results in a negative binomial marginal distribution for Y_i, equivalent to the model referred to as NegBin I by Cameron and Trivedi (1986). The parameter $\phi > 0$ measures the extent of the overdispersion, and the distribution of Y_i converges to a Poisson distribution as $\phi \to 0$.

It should be noted that other types of negative binomial models are also possible. In particular, if η_i is assumed to be Gamma-distributed with constant variance ϕ, the resulting negative binomial distribution has a quadratic mean–variance relationship, $\text{Var}(Y_i) = \mathbb{E}(Y_i)(1+\phi\mathbb{E}(Y_i))$. This model is referred to as NegBin II by Cameron and Trivedi (1986) and various other comparative discussions of the two models have been provided; see for example Hilbe (2011). In Section 5.5 we will present a motivating application in which our additive NegBin I approach is more appropriate, and this model will be our main focus in this paper. However, comparison with the additive NegBin II model will also be considered.

Expressing our model in terms of the usual negative binomial parameters leads to

$$Y_i \sim \text{NegBin}(r_i(\boldsymbol{\theta}), p) \qquad \text{where} \qquad r_i(\boldsymbol{\theta}) = \frac{1}{\phi}N_i\lambda_i = N_i\Lambda(\boldsymbol{x}_i, \boldsymbol{\theta})$$

with $\boldsymbol{\theta} = \boldsymbol{\theta}^*/\phi$ and $p = \phi/(1+\phi)$. By finding the MLEs $\hat{\boldsymbol{\theta}}$ and $\hat{p}$ of this marginal negative binomial distribution, we can transform them back to obtain the MLEs of the coefficient of overdispersion $\hat{\phi}$ and the adjusted rate differences $\hat{\boldsymbol{\theta}}^*$:

$$\hat{\phi} = \frac{\hat{p}}{1-\hat{p}} \qquad \text{and} \qquad \hat{\boldsymbol{\theta}}^* = \hat{\phi}\hat{\boldsymbol{\theta}}. \tag{5.2}$$

However, in practice the additive structure for $r_i(\theta)$ leads to substantial model-fitting challenges stemming from the constraint $r_i(\theta) \geq 0$. Indeed, as discussed in Section 5.6, there is no standard commercial software package that fits this model. In the next section we will describe how the additive structure combined with the fact that p is constant across individuals allows us to use the convolution properties of the negative binomial distribution to construct an underlying latent outcome model. This will then allow us to implement the ECME algorithm (Liu and Rubin, 1994), which provides a stable approach to model fitting.

5.3 Model fitting

In this section we describe how the additive negative binomial model can be reliably fitted using a variant of the ECME algorithm. To do this, we need to first describe the constrained parameter space and a latent outcome model that will be used as the complete-data model in an ECME algorithm. We then describe a variant of the ECME algorithm based on the combinatorial EM algorithm of Marschner (2014).

5.3.1 Parameter space

In describing the model-fitting procedure, it is helpful to keep the categorical and continuous covariates separate. We will examine a model that includes A categorical covariates, $u_{i1}, \ldots, u_{iA}$, and B continuous covariates, $v_{i1}, \ldots, v_{iB}$, for each individual i. Without loss of generality, the possible values of the a^{th} categorical covariate are $\{1, \ldots, k_a\}$, and the b^{th} continuous covariate can take any value in the range $[v_b^{(0)}, v_b^{(1)}]$, where $v_b^{(0)} = \min_i\{v_{ib}\}$ and $v_b^{(1)} = \max_i\{v_{ib}\}$.

The parameter vector $\boldsymbol{\theta}$ associated with this model has $J = 1 + \sum_{a=1}^{A} k_a + B$ components: $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_A, \beta_1, \ldots, \beta_B)^\top$, where α_0 is an intercept term, each $\boldsymbol{\alpha}_a =$

$(\alpha_a(1), \ldots, \alpha_a(k_a))$ is a vector of parameters associated with the a^{th} categorical covariate and β_b is the parameter associated with the b^{th} continuous covariate.

Thus the additive model is

$$\Lambda(\boldsymbol{x}_i, \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta_b v_{ib}, \tag{5.3}$$

which for simplicity we can express in the form equivalent to (5.1)

$$\Lambda(\boldsymbol{x}_i, \boldsymbol{\theta}) = \sum_{j=1}^{J} \theta_j x_{ij},$$

where θ_j is the j^{th} component of the parameter vector $\boldsymbol{\theta}$ and we have defined the covariate vector for individual i as $\boldsymbol{x}_i = (1, \boldsymbol{u}'_{i1}, \ldots, \boldsymbol{u}'_{iA}, v_{i1}, \ldots, v_{iB})$, where $\boldsymbol{u}'_{ia}$ is a vector of length k_a with 1 in the u_{ia}^{th} position and 0 elsewhere.

The parameter space for the additive negative binomial model requires that the expected counts are non-negative and the Gamma scale parameter is positive, that is $(\boldsymbol{\theta}, p) \in \Theta \times \Phi$, where

$$\Theta = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{x}, \boldsymbol{\theta}) \geq 0,\ \boldsymbol{x} \in \mathcal{X}\} \qquad \text{and} \qquad \Phi = (0, 1) \tag{5.4}$$

for a covariate space $\mathcal{X}$, which consists of all possible combinations of the levels of the categorical covariates and any value of the continuous covariates within their observed ranges

$$\mathcal{X} = \prod_{a=1}^{A} \{1, \ldots, k_a\} \times \prod_{b=1}^{B} [v_b^{(0)}, v_b^{(1)}].$$

Note that the parameter space (5.4) does not place any restrictions on the individual components of the parameter vector $\boldsymbol{\theta}$, but we will introduce such restrictions in defining the latent outcome model.

5.3.2 Latent outcome model

In order to define a complete-data model for the ECME algorithm, we begin by considering a particular $\boldsymbol{t} = (t_1, \ldots, t_{A+B})$, giving the reference level for each of the $A + B$ covariates. For each categorical covariate $a = 1, \ldots, A$, we choose $t_a \in \{1, \ldots, k_a\}$ and

set $\alpha_a(t_a) = 0$. For each of the continuous covariates $b = 1, \ldots, B$, the reference level is either the minimum or maximum observed value. That is, $t_{A+b} = v_b^{(\varsigma_b)}$ for a choice of $\varsigma_b \in \{0, 1\}$. For a particular ς_b, we define the shifted covariate $v'_{ib} = (-1)^{\varsigma_b}(v_{ib} - v_b^{(\varsigma_b)})$, which will be non-negative for all i, and (5.3) can be equivalently written as

$$\begin{aligned}\Lambda(\boldsymbol{x}_i, \boldsymbol{\theta}) &= \Lambda(\boldsymbol{x}'_i, \boldsymbol{\theta}') \\ &= \alpha'_0 + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta'_b v'_{ib} \\ &= \sum_{j=1}^{J} \theta'_j x'_{ij},\end{aligned} \tag{5.5}$$

where $\boldsymbol{x}'_i = (1, \boldsymbol{u}'_{i1}, \ldots, \boldsymbol{u}'_{iA}, v'_{i1}, \ldots, v'_{iB})$ and $\boldsymbol{\theta}' = (\alpha'_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_A, \beta'_1, \ldots, \beta'_B)^\top$, with

$$\alpha'_0 = \alpha_0 + \sum_{b=1}^{B} \beta_b v_b^{(\varsigma_b)} \qquad \text{and} \qquad \beta'_b = (-1)^{\varsigma_b}\beta_b. \tag{5.6}$$

For a particular choice of $\boldsymbol{t}$, the complete-data model consists of J independent latent negative binomial random variables with common p underlying the observed Y_i. That is,

$$Y_i = \sum_{j=1}^{J} \mathcal{Y}_i^{(j)},$$

where each $\mathcal{Y}_i^{(j)} \sim \text{NegBin}(r_{ij}(\boldsymbol{\theta}'), p)$ and $r_{ij}(\boldsymbol{\theta}') = N_i \theta'_j x'_{ij}$.

Because the sum of independent negative binomial random variables with common p is itself negative binomial, the complete-data model is equivalent to the observed-data model, with the additional restriction that the $r_{ij}(\boldsymbol{\theta}')$ parameter of each of the latent variable distributions must be non-negative. As each x'_{ij} is defined to be non-negative for all i and j, this natural restriction imposed by the complete-data model for a given $\boldsymbol{t}$ is effectively a non-negativity constraint on each of the components of $\boldsymbol{\theta}'$.

From (5.5) it is clear that for any choice of the reference level vector $\boldsymbol{t}$, $\Lambda(\boldsymbol{t}, \boldsymbol{\theta}) = \alpha'_0$, and the non-negativity constraints on the transformed parameters ensure that the fitted value for any other transformed covariate vector will be larger than this. Thus, the parameter space related to the complete-data model for a particular choice of $\boldsymbol{t}$ is $\Theta(\boldsymbol{t}) \times \Phi$, where

$$\Theta(\boldsymbol{t}) = \{\boldsymbol{\theta} : \Lambda(\boldsymbol{x}, \boldsymbol{\theta}) \geq \Lambda(\boldsymbol{t}, \boldsymbol{\theta}) \geq 0, \, \boldsymbol{x} \in \mathcal{X}\}.$$

There are $\prod_a k_a \times 2^B$ possible choices for the reference level vector $\boldsymbol{t}$, which define a family of complete-data models. Importantly, for any $\boldsymbol{\theta} \in \Theta$, the smallest fitted value must correspond to one of these possible choices of $\boldsymbol{t}$. Hence, if we can find the constrained MLE associated with each of the complete-data models, the MLE for the observed-data model is simply the constrained MLE that attains the highest likelihood. This means that the MLE can be found by cycling through each of the possible choices of $\boldsymbol{t}$, which constitutes an implementation of the combinatorial EM algorithm described by Marschner (2014).

A useful feature of this approach is that it is straightforward to consider a parameter space in which a particular coefficient β_b associated with a continuous covariate is constrained to be non-negative. This can be achieved by considering only the reference level vectors in which $\varsigma_b = 0$. Likewise, a non-positivity constraint can be imposed by considering only the reference level vectors in which $\varsigma_b = 1$. We will take advantage of this feature to include semi-parametric terms in our model, described in Section 5.4.

5.3.3 ECME algorithm

We find the MLE for a particular complete-data model by using the ECME algorithm described by Liu and Rubin (1994). In the ECME algorithm, the M-step of the EM algorithm (Dempster, Laird, and Rubin, 1977) is replaced by a series of conditional maximisation (CM) steps, which act on a subset of the unknown parameters while keeping the others fixed at their current estimates. It differs from the ECM algorithm (Meng and Rubin, 1993) in that some of the CM-steps can be designed to maximise the observed-data log-likelihood rather than the expected complete-data log-likelihood from the E-step. In order to ensure that the likelihood increases monotonically, Meng and Van Dyk (1997) noted that the CM-steps maximising the observed-data log-likelihood must be performed after all steps that maximise the expected complete-data log-likelihood at each iteration.

After removing terms that do not depend on $\boldsymbol{\theta}$ or p, the observed-data log-likelihood for the additive negative binomial model is

$$\ell(\boldsymbol{\theta}, p; \boldsymbol{Y}) = \sum_{i=1}^{n} \log(\Gamma(Y_i + r_i(\boldsymbol{\theta}))) - \log(\Gamma(r_i(\boldsymbol{\theta}))) + r_i(\boldsymbol{\theta}) \log(1-p) + Y_i \log(p),$$

and the complete-data log-likelihood for a chosen $\boldsymbol{t}$ has a similar form:

$$L(\boldsymbol{\theta}', p; \boldsymbol{\mathcal{Y}}) = \sum_{i=1}^{n}\sum_{j=1}^{J} \log(\Gamma(\mathcal{Y}_i^{(j)}+r_{ij}(\boldsymbol{\theta}'))) - \log(\Gamma(r_{ij}(\boldsymbol{\theta}'))) + r_{ij}(\boldsymbol{\theta}')\log(1-p) + \mathcal{Y}_i^{(j)}\log(p).$$

Given a set of starting parameter estimates $\hat{\boldsymbol{\theta}}'_{(0)}$ and $\hat{p}_{(0)}$, the E-step at the $(c+1)^{\text{th}}$ iteration requires calculation of

$$Q(\boldsymbol{\theta}', p \mid \hat{\boldsymbol{\theta}}'_{(c)}, \hat{p}_{(c)}) = \mathbb{E}\left(L(\boldsymbol{\theta}', p; \boldsymbol{\mathcal{Y}}) \mid \boldsymbol{Y}, \hat{\boldsymbol{\theta}}'_{(c)}, \hat{p}_{(c)}\right),$$

for which we use the property that the conditional distribution of negative binomial random variables with common p given their sum is beta-binomial (Wisniewski, 1966). That is,

$$\begin{aligned}\Pr_{(c)}(\mathcal{Y}_i^{(j)} = y \mid Y_i) &= \Pr(\mathcal{Y}_i^{(j)} = y \mid Y_i, \hat{\boldsymbol{\theta}}'_{(c)}, \hat{p}_{(c)}) \\ &= \binom{Y_i}{y} \frac{\mathrm{B}\left(y + r_{ij}(\hat{\boldsymbol{\theta}}'_{(c)}), Y_i + r_i(\hat{\boldsymbol{\theta}}_{(c)}) - (y + r_{ij}(\hat{\boldsymbol{\theta}}'_{(c)}))\right)}{\mathrm{B}\left(r_{ij}(\hat{\boldsymbol{\theta}}'_{(c)}), r_i(\hat{\boldsymbol{\theta}}_{(c)}) - r_{ij}(\hat{\boldsymbol{\theta}}'_{(c)})\right)},\end{aligned}$$

for $y = 0, \ldots, Y_i$, where $\mathrm{B}(\cdot,\cdot)$ denotes the Beta function.

The first CM-step involves conditional maximisation of Q with respect to $\boldsymbol{\theta}'$, holding p constant at its current estimate $\hat{p}_{(c)}$. To do this, for each θ'_j we must find the non-negative root of the derivative

$$\frac{\partial Q}{\partial \theta'_j} = \sum_{i=1}^{n} N_i x'_{ij} \left\{ \sum_{y=0}^{Y_i} \left(\psi(y + r_{ij}(\boldsymbol{\theta}')) - \psi(r_{ij}(\boldsymbol{\theta}'))\right) \Pr_{(c)}(\mathcal{Y}_i^{(j)} = y \mid Y_i) + \log(1 - \hat{p}_{(c)}) \right\}.$$

This does not have an explicit solution, but we can exploit the property of the digamma function

$$\psi(y + r) - \psi(r) = \sum_{k=0}^{y-1} \frac{1}{r+k} \leq \frac{y}{r} \qquad y = 1, 2, \ldots$$

in order to find an upper bound for the root

$$\hat{\theta}'_{j(c+1)} \leq \frac{\sum_{i=1}^{n} \mathbb{E}_{(c)}\left(\mathcal{Y}_i^{(j)} \mid Y_i\right)}{\log(\frac{1}{1-\hat{p}_{(c)}}) \sum_{i=1}^{n} N_i x'_{ij}} = U_{j(c+1)} \qquad \text{where} \qquad \mathbb{E}_{(c)}\left(\mathcal{Y}_i^{(j)} \mid Y_i\right) = Y_i \frac{r_{ij}(\hat{\boldsymbol{\theta}}'_{(c)})}{r_i(\hat{\boldsymbol{\theta}}_{(c)})}.$$

This allows the straightforward use of an omnibus root-finding routine, such as `uniroot`

in R, to find the root in the finite interval $[0, U_{j(c+1)}]$.

In the second CM-step, we keep the additive part $\hat{\boldsymbol{\theta}}'_{(c+1)}$ fixed at its current estimate, and update the estimate for p by maximising the observed-data log-likelihood, giving

$$\hat{p}_{(c+1)} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} r_i(\hat{\boldsymbol{\theta}}_{(c+1)})}.$$

If our starting estimates are within the parameter space for a given complete-data model, the updated estimates at each iteration are guaranteed to remain within the parameter space, and the likelihood will increase until it converges to its maximum value within the restricted parameter space $\Theta(\boldsymbol{t})$.

This process is repeated for each possible choice of the reference vector $\boldsymbol{t}$ in order to find the global maximum, which is the constrained MLE with the highest likelihood. However, there are some situations in which we can stop early, having only searched a subset of the restricted parameter spaces (Marschner, 2014). In particular, if we find a stationary point in the interior of the parameter space $\Theta(\boldsymbol{t})$, we can be sure that it is the overall MLE.

Finally, after finding the overall MLE $(\hat{\boldsymbol{\theta}}', \hat{p})$, we can use the inverse of (5.6) to obtain the unshifted intercept and linear slope parameters, and then (5.2) to calculate the estimated coefficient of overdispersion $\hat{\phi}$, and the estimated rate difference parameters $\hat{\boldsymbol{\theta}}^* = (\hat{\alpha}_0^*, \hat{\boldsymbol{\alpha}}_1^*, \ldots, \hat{\boldsymbol{\alpha}}_A^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_B^*)^\top$.

For categorical covariates, $\hat{\alpha}_a^*(u)$ is the rate difference associated with level u compared to the reference level t_a, adjusted for the other $A + B - 1$ covariates. For continuous covariates, $\hat{\beta}_b^*$ is the adjusted rate difference associated with a one-unit increase in the b^{th} continuous covariate.

5.3.4 Variance estimation

If our distributional assumption and mean model are correct, the MLE resulting from this method will be consistent, with an asymptotic multivariate normal distribution. We can obtain an estimate of the covariance matrix for the parameter estimates $(\hat{\boldsymbol{\theta}}^*, \hat{\phi})$ by using the inverse of the observed information matrix, evaluated at the MLE. The relevant formulae are given in Appendix 5.A.

However, asymptotic normality of the MLE may be questionable if the estimate is on

or close to the boundary of the parameter space. In this case, confidence intervals for parameter estimates may be obtained by using a resampling method such as the bootstrap. The stability of our fitting method ensures that the MLE can be obtained in every resampled dataset, and so there will be no bias caused by non-convergence in some samples.

5.4 Semi-parametric model

In some situations it may be desirable to relax the linearity restriction on the effect of continuous covariates on the expected count or rate, and instead adjust for these covariates semi-parametrically. To allow this, we extend (5.3) to include C additional unspecified functions, so

$$\Lambda(\boldsymbol{x}_i, \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \alpha_a(u_{ia}) + \sum_{b=1}^{B} \beta_b v_{ib} + \sum_{c=1}^{C} f_c(w_{ic}).$$

As demonstrated by Donoghoe and Marschner (2015) for the additive Poisson model, it is possible to incorporate the estimation of the unknown functions f_c into the additive negative binomial model by using B-spline basis functions. This allows us to retain the stability of our estimation method, while still ensuring that the fitted means are non-negative. The approach relies on the fact that it is straightforward to impose non-negativity constraints on any of the parameters associated with continuous covariates. This is easy to do for the additive negative binomial model also, as discussed in Section 5.3.2. We now describe the B-spline model.

5.4.1 B-spline model

For notational convenience, in this subsection we will first consider the case of $C = 1$ and drop the subscript c from the unspecified function f_c. We parameterise these regression functions using the B-spline model

$$f(w) = \sum_{d=1}^{D} \gamma_d B_d(w),$$

where $\{B_d;\, d = 1, \ldots, D\}$ are B-spline basis functions (Ramsay, 1988).

Let $\boldsymbol{\tau}$ be a set of $D+\kappa$ knots on which the B-spline basis is defined, where $\kappa-1$ is the desired degree of the smooth curve and $D-\kappa$ of the knots are distinct turning points in the interior of the range of the continuous covariate. Then the above model means that f is restricted to belong to $\mathcal{B}(\boldsymbol{\tau})$, the function space defined by the B-spline basis on $\boldsymbol{\tau}$. For simplicity, here we use $\kappa = 3$ and choose the $D-3$ turning points to be evenly spaced quantiles of the observed $\{w_i\}$, but these methods will work with any choice of κ and knot vector. We allow the level of smoothness to vary by fitting models with different numbers of turning points and choosing the one with the smallest AIC (Akaike, 1974). Other suitable model selection criteria can also be used, such as the small-sample version, AIC_c (Hurvich, Simonoff, and Tsai, 1998).

Each basis function B_d is positive in (τ_d, τ_{d+3}) and zero elsewhere, meaning that if all of the γ_d coefficients are non-negative, the resulting f will also be strictly non-negative. Furthermore, the B-spline bases are normalised such that $\sum_{d=1}^{D} B_d(w) = 1$ for all w, so we must include an identifiability constraint on the coefficients of each curve.

We begin by choosing a reference level $t_{A+B+1} \in \{1, \ldots, D\}$, and impose an identifiability constraint by setting $\gamma_{t_{A+B+1}} = 0$. The remaining coefficients can be estimated by treating the basis function values $B_d(w_i)$ as continuous covariates in the ECME algorithm, restricting their associated parameters to be non-negative. The resulting f will belong to $\mathcal{B}^+(\boldsymbol{\tau}; t_{A+B+1})$, which denotes the subspace of strictly non-negative curves in $\mathcal{B}(\boldsymbol{\tau})$ that have their shape constrained by the choice of t_{A+B+1}. We repeat this for all D possible choices of t_{A+B+1}, and the estimate with the highest likelihood is the MLE for the semi-parametric model.

Extension to $C > 1$ is straightforward: for each $c = 1, \ldots, C$, we choose a reference level $t_{A+B+c} \in \{1, \ldots, D_c\}$, set $\gamma_{ct_{A+B+c}} = 0$ and use the ECME algorithm to find the MLE, constraining the remaining B-spline coefficients to be non-negative. To find the global maximum, we repeat this for all $\prod_{c=1}^{C} D_c$ possible choices of these reference levels.

5.4.2 Monotonicity restriction

This approach also allows us to impose a monotonicity restriction on any of the smooth curves. A sufficient condition for the monotonicity of f is that coefficients of the B-spline basis functions are themselves strictly non-decreasing, that is $\gamma_1 \leq \cdots \leq \gamma_D$

(Leitenstorfer and Tutz, 2007). If we set $\gamma_1 = 0$ and introduce parameters $\delta_2, \ldots, \delta_D$ that represent the increments between successive coefficients, so

$$\delta_d = \gamma_d - \gamma_{d-1} \qquad d = 2, \ldots, D,$$

an equivalent condition is to require these increments to be non-negative.
Under this parameterisation,

$$f(w) = \sum_{d=2}^{D} \delta_d \left(\sum_{e=d}^{D} B_e(w) \right), \tag{5.7}$$

so a monotonicity restriction can be applied simply by treating the partial sums of B-spline bases in (5.7) as continuous covariates, and constraining their associated parameters to be non-negative in the CEM algorithm. Because we only need to consider one parameterisation for each monotonic regression function, extension to $C > 1$ is trivial, and in fact the inclusion of additional semi-parametric monotonic covariates does not require additional applications of the ECME algorithm.

5.5 Application

The Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study was a double-blind randomised clinical trial examining the effect of fenofibrate treatment on cardiovascular risk in 9795 participants with type II diabetes, aged between 50–75 years over a median follow-up time of 5 years (FIELD Study Investigators, 2005). A pre-specified secondary endpoint of the study was laser photocoagulation treatment for diabetic retinopathy, a microvascular complication of diabetes. Laser therapy is used to slow or prevent vision loss caused by retinopathy, but is associated with visual field reduction and other side-effects (Aiello, 2003), so reducing the need for its use would be a positive outcome.
The vast majority of patients (96%) did not undergo laser therapy during the follow-up period, but individuals could have multiple courses of laser therapy, and we wish to estimate the effect of fenofibrate treatment on the number of laser therapy courses administered per 5 patient-years. In the primary analysis of this data (Keech et al., 2007), a Poisson GLM with a log link function was used to estimate a rate ratio (RR)

for the treatment effect. The overdispersion apparent under the Poisson model was accounted for by rescaling the estimated standard errors.

There was no evidence of an interaction between fenofibrate treatment and the presence of known prior retinopathy at baseline on the multiplicative scale, with an estimated overall RR of 0.63 (95% CI 0.49–0.81). This estimate was applied to the observed mean rates of laser therapy in the placebo group for those with and without known prior retinopathy to obtain estimates of the absolute rate reduction, and hence the expected number of laser therapy courses that may be avoided by administering fenofibrate in these cohorts.

Alternatively, by using an additive model we are able to directly estimate these absolute rate differences and their confidence intervals. The model will also allow us to adjust for individuals' duration of diabetes prior to study entry, which is strongly associated with the risk of retinopathy, and confounded with prior retinopathy status. A negative binomial model for the number of laser therapy events will account for overdispersion with a distributional form.

Assuming no interaction between treatment and known prior retinopathy, we estimated the effect of fenofibrate, first unadjusted for any covariates, then adjusted for baseline retinopathy, in an additive negative binomial model. The results are shown in Figure 5.1 as rate differences per 5 patient-years, along with their 95% confidence intervals.

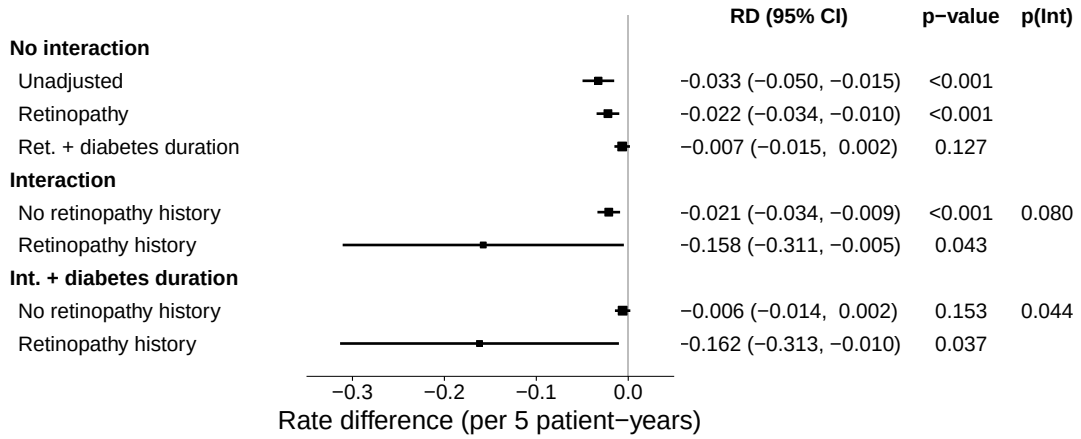


Figure 5.1: Estimated effect of fenofibrate on the rate of laser therapy in the FIELD study from additive negative binomial models, unadjusted and adjusted for prior retinopathy status and diabetes duration. p(Int) denotes the p-value for testing an interaction between treatment and retinopathy history.

Diabetes duration up to 20 years was then entered into the model as a flexible smooth term with a maximum of 5 internal knots. The inclusion of diabetes duration substantially improved the model fit as measured by the AIC (3853.4 with diabetes duration compared to 4014.3 without), and led to a noticeable reduction in the treatment effect estimate.

In order to obtain an estimate of the rate difference for fenofibrate separately in patients with and without known prior retinopathy, the model must include an interaction between treatment allocation and retinopathy status. The estimates from this model, with and without adjustment for diabetes duration, are also shown in Figure 5.1. As expected from the analysis that showed no interaction on a multiplicative scale, the common relative effect of treatment is manifested in a smaller absolute rate difference for individuals at low risk, that is those without known prior retinopathy, compared to those at high risk. After adjustment for diabetes duration, this interaction is marginally significant ($p = 0.044$).

Under this model, the effect of diabetes duration on the rate of laser treatment is assumed to be the same for patients with and without known prior retinopathy. The estimated effect is shown in Figure 5.2, with a 95% confidence interval obtained using the information matrix standard errors. This effect of increasing diabetes duration was allowed to be non-monotonic, and the estimate shows a small reduction in rate just prior to 20 years, which may be due to sampling error. The estimated effect of diabetes duration under a monotonicity constraint is also shown in Figure 5.2.

The estimated rates under our additive NegBin I model were very similar to those under an equivalent additive Poisson model, and did not exceed 0.75 events per 5 patient-years in either case. The coefficient of overdispersion in the negative binomial model was estimated to be $\hat{\phi} = 2.44$ (with estimated standard error 0.25), such that the linear mean–variance relationship has a gradient of 3.44, compared to the equidispersion imposed by the Poisson model. The test statistic for a score test of the Poisson versus the NegBin I distribution, which has a standard normal distribution under the null hypothesis (Cameron and Trivedi, 1986), is 114.6, suggesting strong evidence of overdispersion. Likewise the likelihood ratio test statistic, with a $0.5\chi^2(1)$ distribution under the null (Lawless, 1987), is 1476.9.

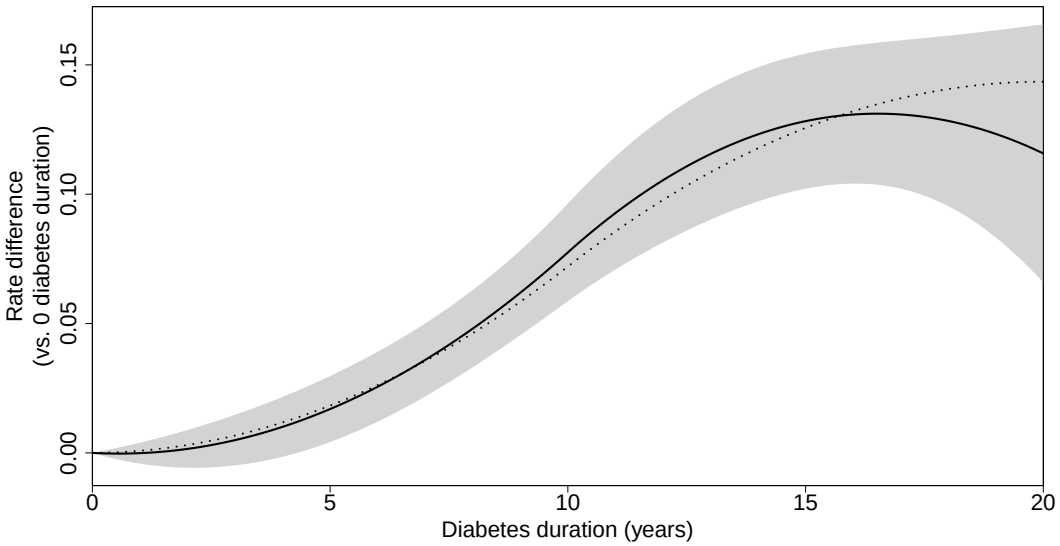
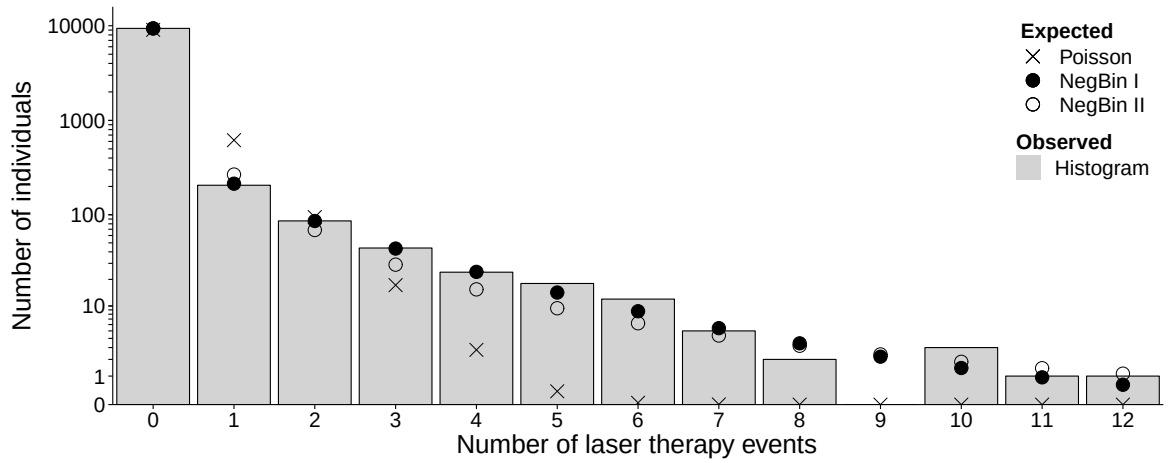


Figure 5.2: Laser therapy rate difference associated with increasing diabetes duration in the FIELD study under an additive negative binomial model, with 95% confidence interval (shaded). The dotted line shows the same fitted rate difference if a monotonicity constraint is applied.

Accounting for overdispersion by using a distributional model allows us to estimate the number of patients in our sample expected to undergo a certain number of laser therapy courses, and these are shown in Figure 5.3, compared to those from the additive Poisson model and the observed histogram. The Poisson model substantially underestimates the number of patients who would not undergo any laser therapy, overestimates the number receiving a single course, and underestimates the number receiving 3 or more courses. In contrast, the NegBin I model shows very good fit to the observed counts, even in the tail of the distribution.

Under an equivalent additive NegBin II model, the mean–variance relationship was estimated to be $\text{Var}(\mu) = \mu + 9.9\mu^2$, and the expected counts under this model are also shown in Figure 5.3. The NegBin II model underestimates the number of patients receiving between 2–6 courses of laser therapy, suggesting that the quadratic mean–variance relationship has inferior fit to the linear overdispersion of the NegBin I model for this data. This is supported by a comparison of AIC values (4132.8 for NegBin II versus 3851.3 for NegBin I) and Chi-squared goodness of fit statistics (48.3 for NegBin II versus 3.44 for NegBin I) for the fitted models. Empirically, a plot of the observed means and variances of the event rate within groups of patients defined by their assigned treatment, prior retinopathy status and diabetes duration (Figure 5.4) also suggests



	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed	9376	206	86	44	24	18	12	5	2	0	3	1	1
Expected													
Poisson	9045.1	618.1	94.4	17.2	2.8	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NegBin I	9372.5	213.9	85.7	43.4	24.2	14.2	8.7	5.4	3.4	2.2	1.4	0.9	0.6
NegBin II	9362.7	267.3	68.7	29.0	15.5	9.4	6.2	4.4	3.2	2.4	1.8	1.4	1.1

FIGURE 5.3: Observed number of individuals with each number of laser therapy courses in the FIELD study, compared to the expected number under additive Poisson, NegBin I and NegBin II models.

that a linear mean–variance relationship is more appropriate for these data.

As an extension, we considered the possibility that the effect of increasing diabetes duration may differ between patients with and without known prior retinopathy. We fitted a similar additive NegBin I model, with a monotonicity constraint but without the assumption of a common effect, and the expected rates under each model are shown in Figure 5.5. The model that allowed different diabetes duration curves had superior fit in terms of AIC (3842.7 versus 3852.8), and shows a dramatic increase in the expected rate between 5 and 15 years of diabetes duration for patients with known prior retinopathy. As can be seen by examining the vertical distance between the pairs of lines in each case however, the estimated treatment effect is virtually identical under both models.

5.6 Discussion

We have described a stable method for finding the MLE in additive negative binomial regression, which allows for estimation of adjusted rate differences in the presence of overdispersion. It also allows for smooth semi-parametric regression using B-splines,

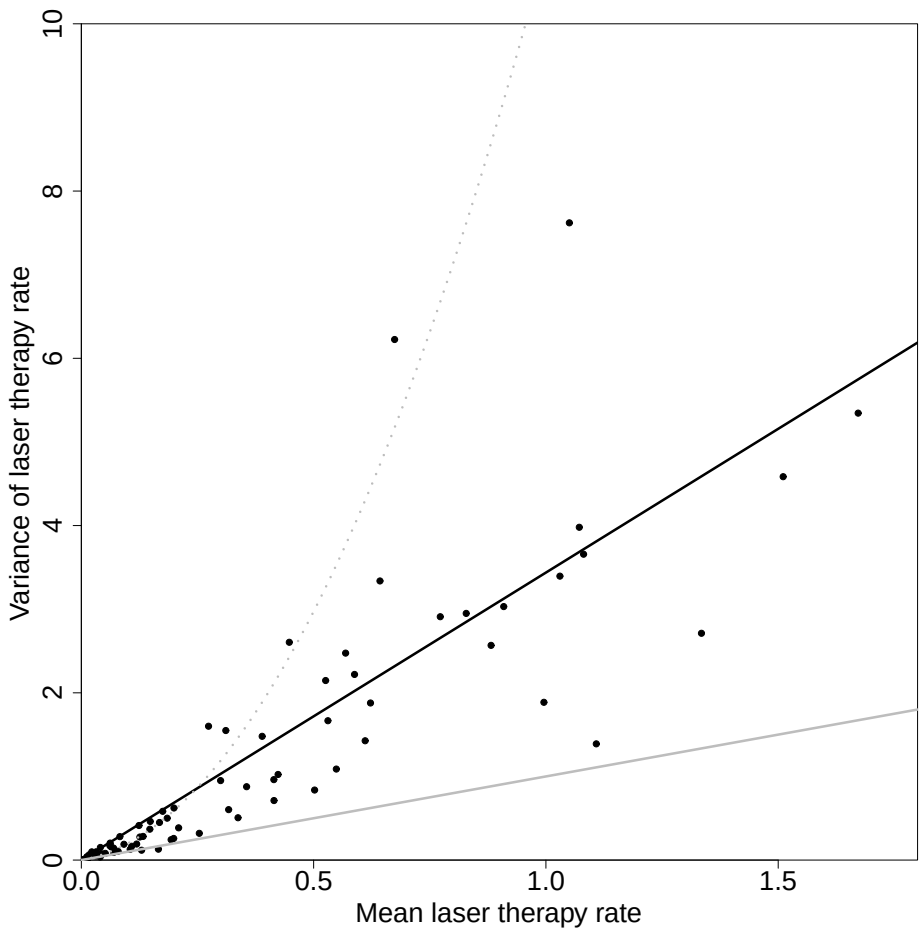


Figure 5.4: Observed mean and variance of laser therapy rates in the FIELD study, for groups defined by assigned treatment, prior retinopathy status and diabetes duration. The solid black line shows the estimated mean–variance relationship from the fitted NegBin I model; the grey lines show the estimated mean–variance relationship under Poisson (solid) and NegBin II (dotted) models.

with optional monotonicity constraints on these curves. The method respects the natural non-negativity restriction on the fitted means, which can cause issues for approaches that employ standard gradient-based algorithms such as Newton–Raphson.

Appendix 5.B gives a detailed summary of existing software that can potentially fit models for estimating adjusted rate differences in the presence of overdispersion. At present, the `gamlss` package in R (Rigby and Stasinopoulos, 2005) is the only other available implementation of a method for fitting additive NegBin I models. However, it employs a variant of the Newton–Raphson algorithm, which may perform poorly when the MLE is on or near the boundary of the parameter space. In the current version (4.3), the routine stops with an error if the estimates move outside the parameter space at any iteration, and although this may be improved by step-size reduction, it is not

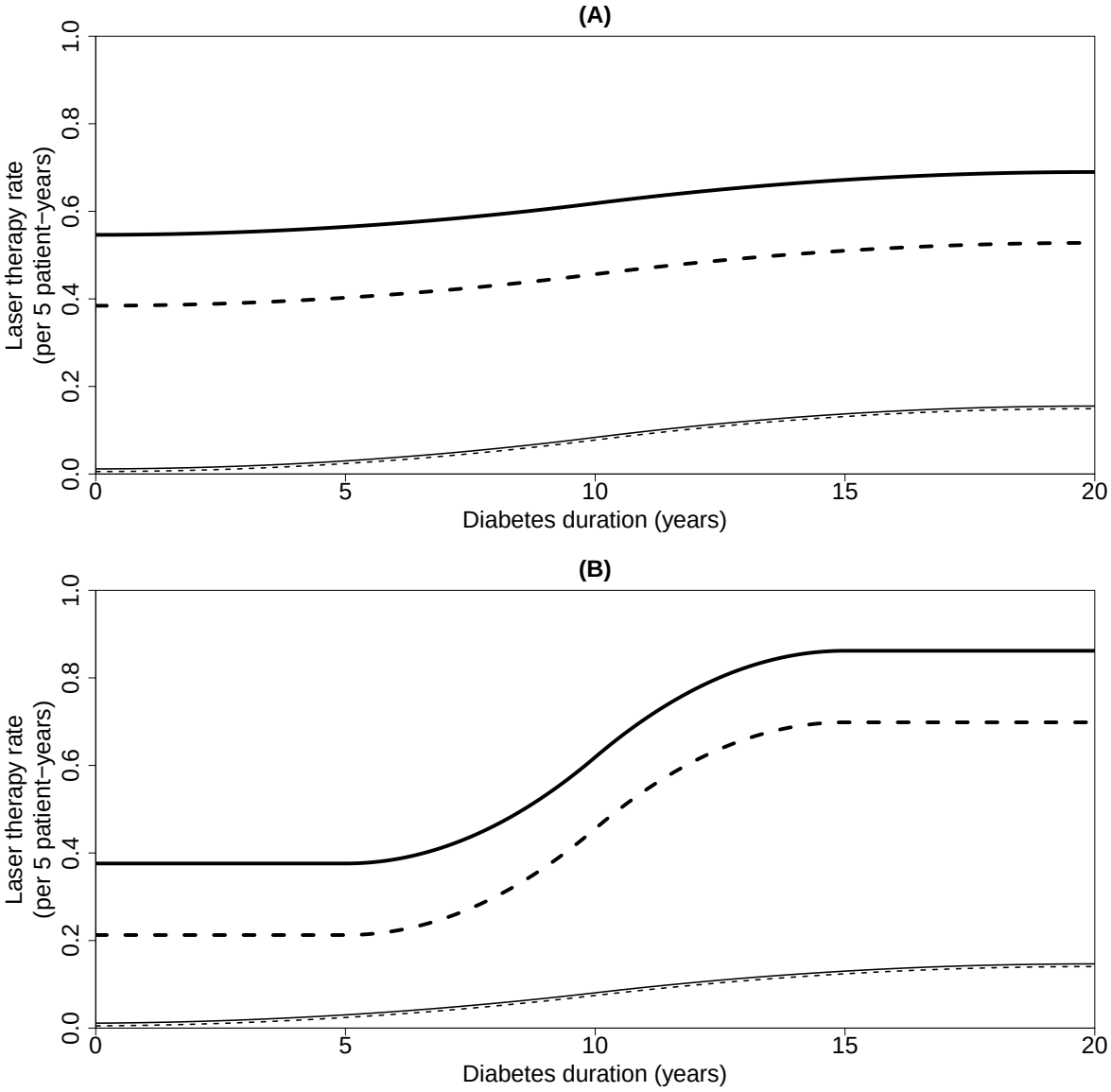


FIGURE 5.5: Fitted rates of laser therapy per 5 patient-years by baseline diabetes duration in the FIELD study under an additive negative binomial model, for patients with (thick lines) and without (thin) known prior retinopathy, assigned to placebo (solid lines) or fenofibrate (dashed) treatment. In (A), the effect of diabetes duration was restricted to be the same for all patients, whereas in (B) it was allowed to differ for those with and without known prior retinopathy.

straightforward to guarantee convergence in all cases. Together with the superiority of the additive NegBin I model in the analysis of Section 5.5, these considerations illustrate the practical usefulness of our method.

Our method has been implemented in the `addreg` package in R (Donoghoe, 2015a), which is available from the Comprehensive R Archive Network (CRAN). Currently the package implements a basic version of the method, searching every possible constrained parameter space consecutively. We plan that future releases will focus on optimising the

computational efficiency using techniques to reduce the number of ECME algorithms that need to be run, such as those discussed by Marschner (2014), and by taking advantage of the independent nature of the multiple ECME algorithms using parallel implementation on a multi-core processor.

The additive NegBin I model that we consider in this paper has a linear relationship between the conditional mean and variance, distinguishing it from the additive NegBin II model, which has a quadratic relationship. Both of these are nested in the NegBin-p model (Cameron and Trivedi, 1986), which could be used to test which is more appropriate, but there are no existing methods for fitting this model with an additive mean. Alternatively, a scatterplot similar to that used by Armitage (1957) or Figure 5.4 can help to distinguish between NegBin I and NegBin II.

Both overdispersion and zero-inflation may be observed in the same data as a result of the data generating process. A possible extension of our method is the inclusion of zero-inflation by introducing a latent Bernoulli random variable into the complete-data model associated with the CEM algorithm. By utilising the methods of Marschner and Gillett (2012) and Donoghoe and Marschner (2014), stability of the algorithm could be maintained for any choice of logit, log or identity link in the regression model for the binary component. Such extensions are the subject of ongoing research.

Software

The methods described in this paper have been implemented in an R package `addreg`, available in the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/package=addreg`.

Supplementary material

Supplementary material for this chapter is presented in Appendices 5.A and 5.B.

Acknowledgments

The authors thank Tony Keech and the FIELD study investigators for providing the laser therapy dataset for analysis. *Conflict of Interest*: None declared.

Funding

This work was supported by The Australian Research Council (DP110101254).

References

Aiello, L. M. (2003). Perspectives on diabetic retinopathy. *American Journal of Ophthalmology* **136**(1): 122–135. DOI: 10.1016/S0002-9394(03)00219-8.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716–723. DOI: 10.1109/tac.1974.1100705.

Armitage, P. (1957). Studies in the variability of pock counts. *The Journal of Hygiene* **55**(4): 564–581. DOI: 10.1017/S0022172400037451.

Bates, G. E. and J. Neyman (1952). Contributions to the theory of accident proneness. II: True or false contagion. *University of California Publications in Statistics* **1**(10): 255–276.

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer-Verlag.

Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* **1**(1): 29–53. DOI: 10.1002/jae.3950010104.

Cameron, A. C. and P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1): 1–38.

Donoghoe, M. W. (2015a). *addreg: Additive Regression for Discrete Data*. R package version 2.0. URL: http://CRAN.R-project.org/package=addreg.

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* **80**: 184–196. DOI: 10.1016/j.csda.2014.06.019.

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* **11**(1): 91–108. DOI: 10.1515/ijb-2014-0044.

FIELD Study Investigators (2005). Effects of long-term fenofibrate therapy on cardiovascular events in 9795 people with type 2 diabetes mellitus (the FIELD study): Randomised controlled trial. *The Lancet* **366**(9500): 1849–1861. DOI: 10.1016/S0140-6736(05)67667-2.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52**(3): 681–700. DOI: 10.2307/1913471.

Greenwood, B. (2005). Interpreting vaccine efficacy. *Clinical Infectious Diseases* **40**(10): 1519–1520. DOI: 10.1086/429833.

Greenwood, M. and G. U. Yule (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* **83**(2): 255–279. DOI: 10.2307/2341080.

Hilbe, J. M. (2011). *Negative Binomial Regression*. 2nd ed. Cambridge: Cambridge University Press.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(2): 271–293. DOI: 10.1111/1467-9868.00125.

Johnson, N. L., S. Kotz, and A. W. Kemp (1993). *Univariate Discrete Distributions*. 2nd ed. New York: Wiley-Interscience: pp. 199–235.

Keech, A. C., P. Mitchell, P. A. Summanen, J. O'Day, T. M. E. Davis, M. S. Moffitt, M. R. Taskinen, R. J. Simes, D. Tse, E. Williamson, A. Merrifield, L. T. Laatikainen, M. C. d'Emden, D. C. Crimet, R. L. O'Connell, and P. G. Colman (2007). Effect of fenofibrate on the need for laser treatment for diabetic retinopathy (FIELD study): A randomised controlled trial. *The Lancet* **370**(9600): 1687–1697. DOI: 10.1016/S0140-6736(07)61607-9.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15**(3): 209–225. DOI: 10.2307/3314912.

Leitenstorfer, F. and G. Tutz (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8**(3): 654–673. DOI: 10.1093/biostatistics/kxl036.

Liu, C. and D. B. Rubin (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**(4): 633–648. DOI: 10.2307/2337067.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**(3): 666–683. DOI: 10.1198/jcgs.2010.09127.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* **24**(6): 921–940. DOI: 10.1007/s11222-013-9411-7.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics* **13**(1): 179–192. DOI: 10.1093/biostatistics/kxr030.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**(2): 267–278. DOI: 10.2307/2337198.

Meng, X.-L. and D. Van Dyk (1997). The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(3): 511–567. DOI: 10.1111/1467-9868.00082.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**(4): 425–441. DOI: 10.1214/ss/1177012761.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**: 507–554. DOI: 10.1111/j.1467-9876.2005.00510.x.

Tang, W., H. He, and X. M. Tu (2012). *Applied Categorical and Count Data Analysis*. Boca Raton, FL: CRC Press: p. 182.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**(3): 439–447. DOI: 10.2307/2334725.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. 5th ed. Berlin: Springer-Verlag.

Winkelmann, R. and K. F. Zimmermann (1992). "Robust Poisson regression". *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling, Munich, 13–17 July 1992*. New York: Springer-Verlag: pp. 201–206.

Wisniewski, T. K. M. (1966). Another statistical solution of a combinatorial problem. *The American Statistician* **20**(3): 25. DOI: 10.1080/00031305.1966.10480403.

Appendix

5.A Information matrix for the additive negative binomial model

Given maximum likelihood estimates $(\hat{\boldsymbol{\theta}}^*, \hat{\phi})$ for the additive NegBin I model defined in Section 5.2, the $(J+1) \times (J+1)$ observed information matrix I_o can be expressed as

$$I_o = \begin{bmatrix} I_1 & I_2 \\ I_2^\top & I_3 \end{bmatrix},$$

where we begin by defining

$$\begin{aligned} \Psi_{0i} &= \psi\left(Y_i + \frac{1}{\hat{\phi}} N_i \Lambda(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^*)\right) - \psi\left(\frac{1}{\hat{\phi}} N_i \Lambda(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^*)\right) \\ &= \psi\left(Y_i + r_i(\hat{\boldsymbol{\theta}})\right) - \psi\left(r_i(\hat{\boldsymbol{\theta}})\right) \\ &= \begin{cases} 0, & Y_i = 0 \\ \sum_{k=0}^{Y_i - 1} (r_i(\hat{\boldsymbol{\theta}}) + k)^{-1}, & Y_i = 1, 2, \ldots \end{cases} \end{aligned}$$

and

$$\begin{aligned} \Psi_{1i} &= \psi_1\left(Y_i + \frac{1}{\hat{\phi}} N_i \Lambda(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^*)\right) - \psi_1\left(\frac{1}{\hat{\phi}} N_i \Lambda(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^*)\right) \\ &= \psi_1\left(Y_i + r_i(\hat{\boldsymbol{\theta}})\right) - \psi_1\left(r_i(\hat{\boldsymbol{\theta}})\right) \\ &= \begin{cases} 0, & Y_i = 0 \\ -\sum_{k=0}^{Y_i - 1} (r_i(\hat{\boldsymbol{\theta}}) + k)^{-2}, & Y_i = 1, 2, \ldots \end{cases} \end{aligned}$$

where ψ_1 denotes the trigamma function. Then I_1 is a symmetric $J \times J$ matrix with (j,k) entry

$$\begin{aligned} I_{1(j,k)} &= -\left.\frac{\partial^2 \ell}{\partial \theta_j^* \partial \theta_k^*}\right|_{(\hat{\boldsymbol{\theta}}^*, \hat{\phi})} \\ &= -\sum_{i=1}^{n} \left(\frac{N_i}{\hat{\phi}}\right)^2 x_{ij} x_{ik} \Psi_{1i}, \end{aligned}$$

I_2 is a $J \times 1$ vector with j^{th} element:

$$\begin{aligned} I_{2(j)} &= -\left.\frac{\partial^2 \ell}{\partial \theta_j^* \partial \phi}\right|_{(\hat{\boldsymbol{\theta}}^*, \hat{\phi})} \\ &= \sum_{i=1}^{n} \frac{N_i x_{ij}}{\hat{\phi}^2} \left(r_i(\hat{\boldsymbol{\theta}}) \Psi_{1i} + \Psi_{0i} + \frac{\hat{\phi}}{\hat{\phi}+1} - \log(\hat{\phi}+1) \right), \end{aligned}$$

and I_3 is a scalar:

$$\begin{aligned} I_3 &= -\left.\frac{\partial^2 \ell}{\partial \phi^2}\right|_{(\hat{\boldsymbol{\theta}}^*, \hat{\phi})} \\ &= \sum_{i=1}^{n} Y_i \frac{2\hat{\phi}+1}{\hat{\phi}^2(\hat{\phi}+1)^2} \\ &\quad - \frac{r_i(\hat{\boldsymbol{\theta}})}{\hat{\phi}} \left\{ \frac{r_i(\hat{\boldsymbol{\theta}})}{\hat{\phi}} \Psi_{1i} + \frac{2}{\hat{\phi}} \left(\Psi_{0i} + \frac{\hat{\phi}}{\hat{\phi}+1} - \log(\hat{\phi}+1) \right) + \frac{\hat{\phi}}{(\hat{\phi}+1)^2} \right\}. \end{aligned}$$

5.B Other approaches for overdispersed count data

In the following sections, we summarise existing approaches that could potentially be used to estimate adjusted rate differences in the presence of overdispersion, and highlight their shortcomings that motivate the study of new methods. We end with some conclusions about the merits of our proposed method compared to these existing approaches.

5.B.1 Negative binomial regression

Table 5.B.1 summarises the availability of fitting methods for additive and multiplicative negative binomial regression models in popular statistical software packages. The majority of these methods use a gradient-based approach such as the Newton–Raphson

algorithm to find the MLE, and this usually performs well when there are no restrictions on the parameter space, as in the case of the multiplicative model. In fact, some texts that discuss negative binomial regression only consider multiplicative models (e.g. Winkelmann, 2008), which is reflected in the fact that some implementations, such as `nbreg` and `gnbreg` in Stata (StataCorp, 2013), only permit a log-linear mean function. Adjusted rate difference estimates can be obtained from a fitted multiplicative model by using the approach of Greenland (2004), but these estimates will be inconsistent if the true mean specification is based on an additive model.

The NegBin II model is often used because it is a member of the linear exponential family if ϕ is fixed, it has a block diagonal information matrix such that the MLEs for the mean and dispersion parameters are independent, and the mean estimates are robust to distributional misspecification (Cameron and Trivedi, 1998, p. 74). Indeed, it is sometimes the only negative binomial distribution considered in texts (e.g. Thurston, Wand, and Wiencke, 2000; Venables and Ripley, 2002) and used by some software packages (e.g. `PROC GENMOD` in SAS; SAS Institute Inc., 2008). But consistency of

Table 5.B.1: Availability of routines for fitting additive and multiplicative negative binomial regression in popular statistical packages.

Software	Function	Additive		Multiplicative	
		NB I	NB II	NB I	NB II
R	`aod::negbin`*				×
	`COUNT::ml.nb1`			×	
	`COUNT::ml.nb2`				×
	`gamlss::gamlss`*†‡	×	×	×	×
	`MASS::glm.nb`		×		×
	`mgcv::gam`†		×		×
	`msme::nbinomial`*		×		×
	`VGAM::vgam`†		×		×
	`VGAM::vglm`		×		×
SAS	`PROC COUNTREG`			×	×
	`PROC GENMOD`		×		×
Stata	`glm`		×		×
	`gnbreg`*			×	×
	`nbreg`			×	×
SPSS	`genlin`		×		×

* These methods allow a regression model for the scale parameter.
† These methods allow smooth semi-parametric terms in the mean regression formula.
‡ In `gamlss`, NegBin I models are accessed with `family = NBII()` and NegBin II models with `family = NBI()`.

the estimated standard errors only holds if the distributional assumption is correct (Cameron and Trivedi, 1998, p. 75), and the superiority of the NegBin I model in the illustrative analysis in the main paper demonstrates the practical usefulness of our NegBin I approach.

Karlis (2001) has proposed an EM-based method for fitting mixed Poisson models, in which the random mixing variables are treated as the missing data. The method is presented for a multiplicative NegBin II model, and requires a log-link Poisson GLM to be fitted in the M-step of each iteration. This could possibly be adapted for an additive model by using an identity-link Poisson GLM at that step, with stability maintained by employing the CEM algorithm of Marschner (2010). It is not clear, however, if this could be easily translated to a NegBin I model.

Other packages that implement general optimisation algorithms can also potentially be used to find the MLE in additive NegBin I models. For example, the `gnlm` package (Lindsey, 2013) only includes the NegBin II distribution as an inbuilt option, but the additive NegBin I likelihood can be entered manually. The `negbin` function in the `aod` package (Lesnoff and Lancelot, 2012) and the fitting functions in the `COUNT` package (Hilbe, 2014) in R assume a multiplicative model, but use the `optim` function, which implements several different maximisation algorithms (R Core Team, 2013). This approach could be adapted to fit an additive NegBin I model. However, these methods do not easily handle the parameter constraints of an additive model, and it may be difficult to ensure convergence close to the boundary.

Without specifying a regression function, a more general mean–variance relationship can be achieved by using the NegBin-p model (Cameron and Trivedi, 1986). Here, the variance of the Gamma-distributed random error is $\phi(N_i\lambda_i)^{p-2}$, resulting in a negative binomial marginal distribution with $\text{Var}(Y_i) = \mathbb{E}(Y_i) + \phi\,[\mathbb{E}(Y_i)]^p$. This nests the NegBin I and NegBin II models when $p = 1$ and $p = 2$ respectively, and this is often noted to be a typical range for p in practice (Engel, 1984). The additional parameter can be estimated using an iterative algorithm such as BHHH (Greene, 2008), and hypothesis tests can be performed to assess for deviation away from NegBin I or NegBin II.

The `gamlss` method allows for an alternative approach to overdispersion modelling within the context of a negative binomial model. This is achieved by allowing the dispersion parameter ϕ, as well as the mean, to vary across individuals through a

regression function. This may be appropriate, for example, if certain subgroups exhibit a greater degree of overdispersion than others.

5.B.2 Alternative models for overdispersion

A number of alternative regression models for dealing with extra-Poisson dispersion have been proposed in the literature. Instead of the Gamma distribution being used for the random effects, Dean, Lawless, and Willmot (1989) use the inverse Gaussian distribution, and Hinde (1982) uses the log-normal distribution. These models can be parameterised so that they share the same mean and variance as the negative binomial distribution, and hence differ only in the higher moments (Winkelmann, 2008, p. 133). Even more generally, Gourieroux, Monfort, and Trognon (1984) considered the situation where the random effect could come from any unspecified exponential family distribution.

Holgate (1970) shows that for any continuous mixing distribution, the resulting marginal likelihood will be unimodal and hence can be maximised by using standard algorithms. For the Poisson–inverse Gaussian model, Guo and Trivedi (2002) define an iterative Broyden–Fletcher–Goldfarb–Shanno algorithm to find the MLE, while Hinde (1982) uses a combination of numerical integration, the EM algorithm and iteratively reweighted least squares for the log-normal mixture. However, in both cases, only multiplicative regression models are defined, and so no consideration is given to the issues that can arise with these algorithms when parameter space restrictions apply.

As with the NegBin II model, the EM algorithm proposed by Karlis (2001) may be used for Poisson–inverse Gaussian models. The step to update the dispersion parameter here has a closed form, so the method appears to be stable for multiplicative models. In an additive model, the M-step of each iteration requires the maximisation of an additive Poisson likelihood, and the CEM algorithm of Marschner (2010) may be used to maintain stability.

The generalised Poisson distribution proposed by Consul (1989) is also an example of a mixed Poisson, although the mixing distribution is difficult to identify (Joe and Zhu, 2005). However, it has the undesirable property that the range of the outcome variable, rather than being unbounded, is dependent on the parameters. This is also true of the generalised event count model (King, 1989; Winkelmann and Zimmermann,

1991), but both distributions have the additional feature of nesting both under- and overdispersion in a single parameter, allowing for hypothesis testing to distinguish between these possibilities.

Less commonly used models include the double Poisson (Efron, 1986), which also allows both under- and overdispersion but only provides approximate expectations, and the Poisson polynomial distribution (Cameron and Johansson, 1997), which has unintuitive parameter interpretations in the regression framework (Winkelmann, 2008, p. 49).

Many of the fitting methods for these models are based on a multiplicative mean specification, and rely on gradient-based algorithms. These could be adapted to additive models, with some step-size adjustment strategy (e.g. Thall, 1988) to help achieve convergence within the parameter space, but we are unaware of any available implementations of such an approach.

In some contexts, it may be appropriate to assume that a particular data generating process is responsible for observed overdispersion. For example, the NegBin_X model of Santos Silva and Windmeijer (2001) results from a stopped-sum characterisation of the negative binomial distribution, and has been used to model the number of visits to a doctor. Similarly, overdispersed data may appear to be zero-inflated, but fitting such a model implies that some of the zero counts are the result of a separate process, which may not be plausible in some scenarios (e.g. Lord, Washington, and Ivan, 2005). Wang, Cockburn, and Puterman (1998) used a combination of EM and quasi-Newton algorithms to fit a finite mixed Poisson model, which is a special case of the finite mixed GLM proposed by McLachlan (1997), for which a multicycle ECM algorithm has been defined. Even more generally, Aitkin (1996) described an EM algorithm to fit a non-parametric finite mixture model, where the mixing distributions themselves are considered to be nuisance functions. Adjusted rate differences could be estimated by such models, although the stability of these methods with an identity link function is unclear.

5.B.3 Model-independent adjustment

A method for estimating an adjusted rate difference and its variance in the presence of overdispersion has been provided by Stukel et al. (1994). This was a generalisation of the approach of Glynn et al. (1993) to allow for an arbitrary mean–variance structure,

and the adjustment is based on direct standardisation of rates. However, this only allows for categorical adjustment covariates, and therefore loses the flexibility of a regression-based approach. Furthermore, this method can be subject to problems with estimation if the strata are sparse (Stijnen and Van Houwelingen, 1993).

A regression-based method was provided by Xu et al. (2010), which allows for any number of adjustment covariates by using an ordinary least squares (OLS) approach on transformed data. Overdispersion is modelled by the introduction of a random frailty term, and robust standard errors are used to account for this. However, because the OLS framework does not impose constraints on the outcome variable, the fitted means resulting from this method are not guaranteed to be non-negative.

5.B.4 Conclusions

In summary, although there is a vast array of potential methods and software for estimating adjusted rate differences in count data subject to overdispersion, none have been designed with the required additive (identity-link) model structure in mind and the consequent constrained maximisation problem that follows. This generally leads to the potential for implementation and stability problems in practice with existing methods. The stability and flexibility of the proposed method, together with the implementation available in the R package `addreg` (Donoghoe, 2015a), is therefore likely to make this a useful method in practice.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**(3): 251–262. DOI: 10.1007/bf00140869.

Cameron, A. C. and P. Johansson (1997). Count data regression using series expansions: With applications. *Journal of Applied Econometrics* **12**(3): 203–223. DOI: 10.1002/(SICI)1099-1255(199705)12:3<203::AID-JAE446>3.0.CO;2-2.

Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* **1**(1): 29–53. DOI: 10.1002/jae.3950010104.

Cameron, A. C. and P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

Consul, P. C. (1989). *Generalized Poisson Distributions: Properties and Applications*. New York: Marcel Dekker Inc.

Dean, C., J. F. Lawless, and G. E. Willmot (1989). A mixed Poisson–inverse-Gaussian regression model. *The Canadian Journal of Statistics* **17**(2): 171–181. DOI: 10.2307/3314846.

Donoghoe, M. W. (2015a). *addreg: Additive Regression for Discrete Data*. R package version 2.0. URL: http://CRAN.R-project.org/package=addreg.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* **81**(395): 709–721. DOI: 10.2307/2289002.

Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statistica Neerlandica* **38**(3): 159–167. DOI: 10.1111/j.1467-9574.1984.tb01107.x.

Glynn, R. J., T. A. Stukel, S. M. Sharp, T. A. Bubolz, J. L. Freeman, and E. S. Fisher (1993). Estimating the variance of standardized rates of recurrent events, with application to hospitalizations among the elderly in New England. *American Journal of Epidemiology* **137**(7): 776–786.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52**(3): 681–700. DOI: 10.2307/1913471.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters* **99**(3): 585–590. DOI: 10.1016/j.econlet.2007.10.015.

Greenland, S. (2004). Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology* **160**(4): 301–305. DOI: 10.1093/aje/kwh221.

Guo, J. Q. and P. K. Trivedi (2002). Flexible parametric models for long-tailed patent count distributions. *Oxford Bulletin of Economics and Statistics* **64**(1): 63–82. DOI: 10.1111/1468-0084.00004.

Hilbe, J. M. (2014). *COUNT: Functions, data and code for count data*. R package version 1.3.2. URL: http://CRAN.R-project.org/package=COUNT.

Hinde, J. (1982). "Compound Poisson regression models". *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*. **14**. New York: Springer-Verlag. Chap. 11: pp. 109–121.

Holgate, P. (1970). The modality of some compound Poisson distributions. *Biometrika* **57**(3): 666–667. DOI: 10.2307/2334787.

Joe, H. and R. Zhu (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal* **47**(2): 219–229. DOI: 10.1002/bimj.200410102.

Karlis, D. (2001). A general EM approach for maximum likelihood estimation in mixed Poisson regression models. *Statistical Modelling* **1**(4): 305–318.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science* **33**(3): 762–784. DOI: 10.2307/2111071.

Lesnoff, M. and R. Lancelot (2012). *aod: Analysis of Overdispersed Data*. R package version 1.3. URL: http://CRAN.R-project.org/package=aod.

Lindsey, J. (2013). *gnlm: Generalized Nonlinear Regression Models*. R package version 1.0. URL: popgen.unimaas.nl/~jlindsey/rcode.html.

Lord, D., S. P. Washington, and J. N. Ivan (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention* **37**(1): 35–46. DOI: 10.1016/j.aap.2004.02.004.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**(3): 666–683. DOI: 10.1198/jcgs.2010.09127.

McLachlan, G. J. (1997). On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research* **6**(1): 76–98.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Santos Silva, J. M. C. and F. Windmeijer (2001). Two-part multiple spell models for health care demand. *Journal of Econometrics* **104**(1): 67–89. DOI: 10.1016/S0304-4076(01)00059-8.

SAS Institute Inc. (2008). *SAS/STAT Software, Version 9.2*. Cary, NC. URL: http://www.sas.com/.

StataCorp (2013). *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.

Stijnen, T. and H. C. Van Houwelingen (1993). Relative risk, risk difference and rate difference models for sparse stratified data: A pseudo likelihood approach. *Statistics in Medicine* **12**(24): 2285–2303. DOI: 10.1002/sim.4780122406.

Stukel, T. A., R. J. Glynn, E. S. Fisher, S. M. Sharp, G. Lu-Yao, and J. E. Wennberg (1994). Standardized rates of recurrent outcomes. *Statistics in Medicine* **13**(17): 1781–1791. DOI: 10.1002/sim.4780131709.

Thall, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data. *Biometrics* **44**(1): 197–209. DOI: 10.2307/2531907.

Thurston, S. W., M. P. Wand, and J. K. Wiencke (2000). Negative binomial additive models. *Biometrics* **56**(1): 139–144. DOI: 10.1111/j.0006-341X.2000.00139.x.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. 4th ed. New York: Springer: pp. 206–208.

Wang, P., I. M. Cockburn, and M. L. Puterman (1998). Analysis of patent data: A mixed-Poisson-regression-model approach. *Journal of Business & Economic Statistics* **16**(1): 27–41. DOI: 10.2307/1392013.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. 5th ed. Berlin: Springer-Verlag.

Winkelmann, R. and K. F. Zimmermann (1991). A new approach for modeling economic count data. *Economics Letters* **37**(2): 139–143. DOI: 10.1016/0165-1765(91)90122-2.

Xu, Y., Y. B. Cheung, K. F. Lam, S. H. Tan, and P. Milligan (2010). A simple approach to the estimation of incidence rate difference. *American Journal of Epidemiology* **172**(3): 334–343. DOI: 10.1093/aje/kwq099.

6

Discussion

In this thesis, we have developed new methodology for fitting semi-parametric regression models for risks and rates. These methods allow us to estimate adjusted risk differences, rate differences and relative risks, which are useful measures of effect in clinical studies. In this final chapter, we provide an overview of the methods, their implementation and potential future extensions.

6.1 Overview

In Chapter 3 we defined a new algorithm for maximum likelihood estimation in additive (identity-link) binomial generalised linear models (GLMs), and demonstrated its reliability and flexibility compared to existing methods. We then extended this method in Chapter 4 to allow for flexible smooth semi-parametric terms in a generalised additive model (GAM), as well as adapting existing stable methods for additive Poisson and log-link binomial models in a similar way. Finally, in Chapter 5 we defined an algorithm

for maximum likelihood estimation in additive negative binomial regression models, providing a method for estimating semi-parametrically adjusted rate differences in the presence of overdispersion.

Our methods are particularly useful when a standard GLM routine, such as the Newton–Raphson or Fisher scoring algorithms that are implemented in most popular statistical software, fails to converge to the maximum likelihood estimate (MLE). As described in Section 2.1.4, these algorithms do not have a natural way of dealing with the parameter constraints that are required in many non-canonical models to ensure that the fitted risks and rates take valid values. In an application to real data presented in Section 3.3.3, we showed an example in which the `glm` function in R could not converge to the MLE, with the parameter estimates becoming caught in a periodic cycle. In simulations presented in Section 3.4, we saw that SAS's `GENMOD` procedure suffered convergence issues with additive binomial models, particularly when the MLE was near the boundary of the parameter space.

Semi-parametric models provide additional flexibility for situations in which we do not wish to specify a parametric functional form for the relationship between a covariate and the risk or rate. We showed an example of this in Section 4.5. Methods for semi-parametric regression models typically rely on the same gradient-based algorithms as GLMs for fitting the parametric components, and can suffer similar convergence problems. In particular, `PROC GAM` in SAS only permits models that use the canonical link function, while `gamlss` in R is the only other available method for fitting additive NegBin I models, but has no inbuilt method for ensuring that the necessary parameter constraints are enforced.

Each of the methods presented in this thesis is an application of a combinatorial EM algorithm. These algorithms provide stable convergence to the MLE and ensure that the parameter estimates always remain within the parameter space at every iteration. Even when a standard approach successfully converges to the MLE, if it is on or close to the boundary of the parameter space, we cannot rely on asymptotic normality to estimate confidence intervals for the model parameters. In such situations, we can instead use bootstrapping, and as demonstrated in Sections 3.3.3 and 4.5, the stability of our approach is important because no bias will be introduced by conditioning on convergence to the MLE in bootstrapped samples.

The models we have examined in this thesis are not restricted to clinical settings: risk differences, rate differences and relative risks are relevant outside of biostatistics. In epidemiology, prevalence can be viewed as the probability that an individual in the population under study has the disease at the observation time, and so additive and multiplicative binomial models can be used to estimate prevalence differences and ratios, both of which are important effect measures (Pearce, 2004). In an epidemiological cohort study, if observed over a fixed time period for all individuals, incidence is also a measure of risk, but may alternatively be expressed as a rate (Rothman and Greenland, 1998). Attributable risk is the absolute difference in incidence (Koepsell and Weiss, 2003), and thus can be estimated by using an additive binomial, Poisson or negative binomial model.

In Section 3.6, we gave some examples of recent applications of additive binomial models in areas such as econometrics and psychometrics. In count data, overdispersion has been observed in a wide range of applications; for example, ecological species counts (Ver Hoef and Boveng, 2007), patent issuance (Hausman, Hall, and Griliches, 1984), violent incidents in behavioural science (Gardner, Mulvey, and Shaw, 1995), and motor vehicle accidents (Lord and Mannering, 2010). In such situations, additive negative binomial models allow estimation of adjusted absolute differences in expected counts or rates, which may be desired due to their simple interpretation, or may provide superior fit over a multiplicative model. Semi-parametric regression of binary and count outcomes has been used for descriptive and exploratory analyses in studies of bank mergers (Behr and Heid, 2011), landslide susceptibility (Goetz, Guthrie, and Brenning, 2011) and social unrest (Yeeles, 2015), and the stable methods described in this thesis for non-canonical models are similarly applicable in such areas. Thus, while the methodology developed in this thesis has been illustrated using clinical applications, it is likely to be relevant for a broad range of other areas.

6.2 Software

The algorithms described in this thesis have been implemented in two open-source software packages for the R computing environment: `logbin` and `addreg`. Both of these packages are available online at the Comprehensive R Archive Network (CRAN).

The full documentation for these packages is reproduced in Appendices A and B, which includes example code for running the main functions. We provide a brief overview of each package here.

6.2.1 `logbin`

The `logbin` package implements the CEM algorithm for log-binomial regression models described by Marschner and Gillett (2012) and extended to include semi-parametric regression in Chapter 4 of this thesis.

The workhorse function that performs the constrained maximum likelihood estimation for a particular parameterisation, `nplbin`, is based on code published in the supplementary materials of Marschner and Gillett (2012). The main function `logbin`, and the auxiliary functions that perform the necessary data manipulations, however, have been rewritten so as to work with the `formula` and `model.frame` structures that are used by the `glm` function in R. In particular, each categorical covariate must be entered into the model as a `factor`, distinguishing them from continuous covariates and allowing the possible reference levels to be easily identified.

The value returned by `logbin` is of class `logbin`, a subclass of `glm` and `lm`. This means that many of the functions associated with these classes — for example, `summary`, `vcov` and `anova` — are also compatible with results from a call to `logbin`.

Semi-parametric regression is performed by using the `logbin.smooth` function, and including one or more `B()` (B-spline) or `Iso()` (isotonic step function) terms in the model formula. Workhorse functions identify the possible reference levels, and iterate through these by creating the associated basis functions and passing the appropriate call to `logbin`. For terms that use a B-spline basis, the set of internal turning points can be specified by the user through the `knots` argument, or a range for the possible number of internal knots can be set in `knot.range`. In the latter case, for each number in the range, the knots are placed at evenly-spaced quantiles of the covariate, and the number of knots that gives the smallest AIC_c is chosen for the resulting model.

The `mono` argument may be used to reduce the number of times that the EM algorithm must be applied, if we are certain that the relationship between a particular covariate and risk is monotonically non-decreasing. For continuous covariates, it constrains

the associated parameter to be non-negative by only considering one of the two possible reference levels. For semi-parametric terms, the fitted curve is constrained to be monotonic by the method described in Section 4.4.5, which also only requires a single parameterisation for that covariate.

The `start` argument allows the user to specify a starting value for the parameter estimates in the EM algorithm, which may also improve the speed of convergence, but only if the MLE is a stationary point in the interior of the parameter subspace in which the `start` value lies. If not, the constrained MLE will be a point on the boundary of the parameter subspace, and the algorithm will continue searching for the global MLE in the other parameter subspaces.

Further documentation on the `logbin` package is provided in Appendix A, along with some examples using datasets available in the `glm2` package (Marschner, 2014), including the ASSENT-2 study discussed in Chapters 3 and 4.

6.2.2 `addreg`

The `addreg` package fits additive Poisson (Marschner, 2010) and additive binomial (Chapter 3) models, including semi-parametric regression (Chapter 4), as well as additive negative binomial regression models (Chapter 5).

It is structured similarly to the `logbin` package, with a single main function, `addreg`, which calls the appropriate workhorse functions to manipulate the data and fit the model, depending on the specified `family`. The `nnpois` function, which implements the EM algorithm for a constrained additive Poisson model, is based on code published in the supplementary material of Marschner, Gillett, and O'Connell (2012). Additive binomial models (`family = binomial`) are fit by manipulating the data as described in Chapter 3 and making a second call to `addreg` with `family = poisson`.

The `nnnegbin` workhorse function is similar to `nnpois`, except that it implements an ECME algorithm in order to estimate the rate difference and overdispersion parameters simultaneously. In this case, the convergence criterion (2.3) may be applied in different ways: in testing, we found the best results when convergence was declared if the condition was met separately for both the vector of rate difference parameters and the coefficient of overdispersion.

For negative binomial models, it is important to note that there is no unique saturated

model, and so the deviance returned by `addreg` for a particular fitted model is relative to a saturated model with the same estimated overdispersion parameter. It is not appropriate to use likelihood ratio tests for comparing negative binomial models, as the dispersion will not be the same in each, and hence they are not nested. Another model selection criterion such as AIC or AIC_c should be used instead.

Further documentation on the `addreg` package is available in Appendix B, along with examples using datasets available in the `glm2` package (Marschner, 2014).

6.3 Future directions

Combinatorial EM algorithms are often useful in constrained maximum likelihood estimation problems where the observed outcome variable can be represented as a function of a collection of unobserved latent outcome variables. The applications in this thesis suggest a range of more complex models for which a CEM algorithm may be applied in order to achieve a stable fit.

For example, in Chapter 3, we used a special case of the multinomial–Poisson transformation in order to fit an additive binomial model for estimating adjusted risk differences. With some modifications to ensure that the parameter space constraints are respected, this could be extended to be used in a more general additive multinomial model for ordinal or nominal outcome data. In such a model, the parameter estimates would correspond to the absolute differences in the probability of the outcome taking a particular value.

Alternatively, the additive binomial model could be extended by combining the approach with a similar idea to that used by Marschner and Gillett (2012) for relative risk regression. There, the binary outcome is expressed as the product of a collection of latent binary outcome variables, and an additive model underlying each of these would result in a stratified additive binomial model, similar to the stratified additive Poisson model of Marschner, Gillett, and O'Connell (2012). Such a model could be used in situations in which the risk depends multiplicatively on some covariates, but additively on others.

The stratified additive Poisson model could also be extended to account for overdispersion by instead using a negative binomial model for the outcome and adapting the

algorithm appropriately, using the ideas for additive negative binomial models discussed in Chapter 5.

Zero-inflated additive Poisson and negative binomial models could also be fitted using a CEM algorithm. Under such a model, the outcome is zero with some probability, and otherwise comes from the proposed count distribution. A latent binary random variable could be used to model this mixing, using a logistic, multiplicative or additive structure.

Each of these extensions could incorporate semi-parametric regression by using B-splines as in Chapter 4. The semi-parametric regression itself could be extended to allow for additional flexibility through the inclusion of penalised maximum likelihood estimation, and this can be achieved while maintaining stability, as discussed briefly in Section 4.7. However, such extensions are likely to be quite computationally expensive, and may only become viable with greater processor speed and large-scale parallelism.

Overall, the methods described in this thesis have a wide range of applications both within and outside biostatistics. Extensions to these approaches could provide additional flexibility for risk and rate modelling, while maintaining the stability of the procedures presented in this thesis.

A

`logbin` package documentation

This appendix contains the documentation for the R package `logbin`, which implements methods for fitting log-link binomial GLMs and GAMs, as described in Marschner and Gillett (2012) and Chapter 4 of this thesis.

The package is available online from the Comprehensive R Archive Network:

Donoghoe, M. W. (2015b). *logbin: Relative Risk Regression Using the Log-Binomial Model*. R package version 1.2. URL: http://CRAN.R-project.org/package=logbin.

The function `nplbin`, which implements the constrained log-binomial regression EM algorithm, is based on code written by Alexandra Gillett and published as supplementary material to Marschner and Gillett (2012). All other functions were written entirely by the candidate.

Package 'logbin'

Title Relative Risk Regression Using the Log-Binomial Model

Description Methods for fitting log-link GLMs and GAMs to binomial data, using EM-type algorithms with more stable convergence properties than standard methods.

Version 1.2

Date 2015-05-12

Author Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

Maintainer Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

Depends R ($\geq$ 3.0.1)

Imports splines, glm2

License GPL ($\geq$ 2)

LazyData true

`logbin-package` *Relative Risk Regression Using the Log-Binomial Model*

Description

Methods for fitting log-link GLMs and GAMs to binomial data, using EM-type algorithms with more stable convergence properties than standard methods.

Details

Package:	logbin
Type:	Package
Version:	1.2
Date:	2015-05-12
License:	GPL ($\geq$ 2)

This package provides methods to fit generalised linear models (GLMs) and generalised additive models (GAMs) with log link functions to binomial data. It has two primary functions: `logbin` and `logbin.smooth`, together with various supporting functions.

It is useful in two main situations. The first is when a standard GLM routine, such as `glm`, fails to converge with such a model. The second is when a flexible semi-parametric component is desired in these models. One of the main purposes of this package is to provide parametric and semi-parametric adjustment of relative risks.

The computational method is a combinatorial EM algorithm (Marschner, 2014), which accommodates the parameter constraints and is more stable than iteratively reweighted least squares. A collection of restricted parameter spaces is defined which covers the full parameter space, and the EM algorithm is applied within each restricted parameter space in order to find a collection of restricted maxima of the log-likelihood function, from which can be obtained the global maximum over the full parameter space.

The methodology implemented in this package is presented in Marschner and Gillett (2012) and Donoghoe and Marschner (2015).

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

Maintainer: Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

References

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics*

11(1): 91–108.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* 13(1): 179–192.

See Also

`glm`

Examples

```
## For examples, see example(logbin) and example(logbin.smooth)
```

`anova.logbin` *Analysis of Deviance for logbin Fits*

Description

Compute an analysis of deviance table for more than one GLM fitted using `logbin`.

Usage

```
## S3 method for class 'logbin'
anova(object, ..., test = NULL)
```

Arguments

`object, ...`	objects of class `"logbin"`, typically the result of a call to `logbin`, or a `list` of `objects` for the "`logbinlist`" method.
`test`	a character string, (partially) matching one of `"Chisq"`, `"LRT"`, `"Rao"`, `"F"` or `"Cp"`. See `stat.anova`.

Details

Unlike `anova.glm`, specifying a single object is not allowed.

The table has a row for the residual degrees of freedom and deviance for each model. For all but the first model, the change in degrees of freedom and deviance is also given. (This only makes statistical sense if the models are nested.) It is conventional to list the models from smallest to largest, but this is up to the user.

Models where the MLE lies on the boundary of the parameter space will be automatically removed from the list (with a warning), because asymptotic results to not apply to such models.

The table will optionally contain test statistics (and p-values) comparing the reduction in deviance for the row to the residuals. Mallows' *Cp* statistic is the residual deviance plus twice the estimate of σ^2 times the residual degrees of freedom, which

is closely related to AIC. You can also choose "LRT" and "Rao" for likelihood ratio tests and Rao's efficient score test. The former is synonymous with "Chisq" (although both have an asymptotic chi-square distribution).

Value

An object of class "anova" inheriting from class "data.frame".

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

See Also

logbin, anova.glm, anova

Examples

```
## For an example, see example(logbin)
```

B.Iso *Defining Smooths in logbin.smooth Formulae*

Description

Function used in the definition of smooth terms within logbin.smooth model formulae. The function does not evaluate a smooth — it exists purely to help set up a model using smooths.

Usage

```
B(..., knots = NULL, knot.range = 0:5)

Iso(...)
```

Arguments

...	variable that this smooth is a function of. Note that unlike gam, smooths that are functions of more than one variable are not supported.
knots	*unique* positions of *interior* knots of a B-spline basis. Boundary knots are created automatically.
knot.range	if knots is not specified, a vector containing a series of non-negative integers denoting the number of *interior* knots for which the model will be fit. These are placed at evenly-spaced quantiles of the observed covariate values. At least one of knots or knot.range must be non-missing.

Details

The function does not evaluate the variable arguments; the output from this function is passed as part of the arguments to `logbin.smooth.design`, which constructs the actual basis functions.

`B` is used to specify an order-3 B-spline basis (which can be restricted to be monotonically non-decreasing via the `mono` argument in `logbin.smooth`). If `length(knot.range) > 1`, models with each of the specified number of interior knots will be fit, and the model with the best (smallest) `aic.c` will be returned.

`Iso` is used to specify an isotonic basis, designed such that the resulting function has non-negative increments at each observed covariate value. When `Iso` is used, the resulting function will always be monotonically non-decreasing, regardless of the value of `mono`.

Value

An object of class `"B.smooth"` (for `B`) or `"Iso.smooth"` (for `Iso`), which is a list with the following elements:

`term`	name of the term provided in the `...` argument.
`termlabel`	label for the term in the model; e.g. for term `"x"` it will be `"B(x)"` or `"Iso(x)"`.
`knots`	vector of interior knots (if specified). `NA` for `Iso`.
`knot.range`	vector of number of interior knots. `NA` for `Iso`.

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

See Also

`logbin.smooth`, `logbin.smooth.design`

`s` performs a similar function in the `mgcv` package.

Examples

```
## See example(logbin.smooth) for an example of specifying smooths in
##  model formulae.
```

`confint.logbin` *Confidence Intervals for logbin Model Parameters*

Description

Computes confidence intervals for one or more parameters in a fitted `logbin` model.

Usage

```
## S3 method for class 'logbin'
confint(object, parm, level = 0.95, ...)
```

Arguments

`object`	a fitted model object, resulting from a call to `logbin`.
`parm`	a specification of which parameters are to be given confidence intervals, either a vector of numbers or a vector of names. If missing, all parameters are considered.
`level`	the confidence level required.
`...`	additional argument(s) passed to `confint.default`.

Details

Calculates confidence intervals for model parameters assuming asymptotic normality and using the result from `vcov.logbin(object)`. As such, if the MLE is on the boundary of the parameter space, (as per `object$boundary`) the normality assumption is invalid and `NA` is returned.

Value

A matrix (or vector) with columns giving lower and upper confidence limits for each parameter. These will be labelled as (1-level)/2 and 1-(1-level)/2 in % (by default 2.5% and 97.5%).

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

See Also

`confint.default`, `vcov.logbin`

Examples

```
## For an example, see example(logbin)
```

`contr.isotonic.rev` *Contrast Matrix for Reversed Isotonic Covariate*

Description

Return something similar to a contrast matrix for a categorical covariate that we wish to be monotonically non-decreasing in a specified order.

Usage

```
contr.isotonic.rev(n, perm, contrasts = TRUE, sparse = FALSE)
```

Arguments

`n`	a vector of levels for a factor, or the number of levels.
`perm`	a permutation of the levels of `n` (or of the numbers `1:n`), which define the order in which the coefficients must be monotonically non-decreasing.
`contrasts`	a logical indicating whether constrasts should be computed.
`sparse`	included for compatibility reasons. Has no effect.

Details

This function is used within `logbin.design` for categorical covariates with a specified order under a particular parameterisation. This is required if a categorical covariate is defined as monotonic.

In the order specified by `perm`, the coefficient associated with each level is the sum of increments between the following levels. That is, if there are a total of k levels, the first level is defined as $d_2 + d_3 + d_4 + \cdots + d_k$, the second as $d_3 + d_4 + \cdots + d_k$, the third as $d_4 + \cdots + d_k$, and so on. In fitting the model, these increments are constrained to be non-positive.

Note that these are not 'contrasts' as defined in the theory for linear models, rather this is used to define the `contrasts` attribute of each variable so that `model.matrix` produces the desired design matrix.

Value

A matrix with `n` rows and k columns, with $k = n - 1$ if `contrasts` is `TRUE` and $k = n$ if `contrasts` is `FALSE`.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

See Also

logbin.design, which uses contr.isotonic.rev to create the design matrix using model.matrix.

contr.treatment, contrasts for their usual use in regression models.

Examples

```
contr.isotonic.rev(4,1:4)
contr.isotonic.rev(4,c(1,3,2,4))

# Show how contr.isotonic.rev applies within model.matrix
x <- factor(round(runif(20,0,2)))
mf <- model.frame(~x)
contrasts(x) <- contr.isotonic.rev(levels(x), levels(x))
model.matrix(mf)
```

---

conv.test *Convergence Test Based on L2 Norm*

Description

Performs a test of convergence based on the L2 norm of the change in the parameter estimates.

Usage

```
conv.test(theta1, theta2, epsilon)
```

Arguments

theta1	vector of parameter estimates at previous step.
theta2	vector of parameter estimates at current step.
epsilon	positive convergence tolerance.

Details

This is used as the convergence test in the logbin fitting functions, because the EM algorithm may converge slowly such that the test based on the deviance used in glm.fit (see glm.control) may report convergence at a point away from the actual optimum.

Value

A logical; TRUE if sqrt(sum((theta1-theta2)**2))/sqrt(sum(theta1**2)) < epsilon; FALSE otherwise.

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

Examples

```
theta.old <- c(-4,-5,-6)
theta.new <- c(-4.05,-5,-6)

conv.test(theta.old, theta.new, 0.01)
conv.test(theta.old, theta.new, 0.005)
```

---

interpret.logbin.smooth *Interpret a logbin.smooth Formula*

Description

This is an internal function of package logbin. It is a service routine for logbin.smooth which interprets the smooth parts of the model formula and returns modified formulas to be used in the fitting functions.

Not normally called directly.

Usage

```
interpret.logbin.smooth(formula)
```

Arguments

formula	A formula as supplied to logbin.smooth, which includes at least one B or Iso term.

Value

A list with components:

full.formula	a formula object which is the same as the formula supplied, but with additional arguments removed from the smooth terms. E.g. B(x, knot.range = 0:2) would appear as B(x) in this formula.
fake.formula	a formula object which is the same as the formula supplied, but with smooth terms replaced by their covariates alone. E.g. B(x, knot.range = 0:2) would appear as x in this formula. Used to construct the model matrix.
smooth.spec	a named list containing the results of evaluating the smooth terms. See B and Iso for details.
smooth.ind	a vector containing the indices of the smooth components in the formula.
terms	the result of running terms.formula(formula, specials = c("B", "Iso")).

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

See Also

logbin.smooth

Examples

```
# Specify a smooth model with knot.range
res <- interpret.logbin.smooth(y ~ B(x, knot.range = 0:2) + x2)
# The knot.range is removed from the full.formula...
print(res$full.formula)
# ...but is stored in the $smooth.spec component of the result:
print(res$smooth.spec$x$knot.range)
```

---

logbin *Log-Binomial Regression*

Description

logbin fits relative risk (log-link) binomial regression models using a stable combinatorial EM algorithm.

Usage

```
logbin(formula, mono = NULL, data, subset, na.action, start = NULL,
       offset, control = list(...), model = TRUE, warn = TRUE, ...)
```

Arguments

formula	an object of class "formula" (or one that can be coerced into that class): a symbolic description of the model to be fitted. The details of model specification are given under "Details". Note that the model must contain an intercept, and 2nd-order terms (such as interactions) or above are currently not supported — see "Note".
mono	a vector indicating which terms in formula should be restricted to have a monotonically non-decreasing relationship with the outcome. May be specified as names or indices of the terms.
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which logbin is called.
subset	an optional vector specifying a subset of observations to be used in the fitting process.

na.action	a function which indicates what should happen when the data contain NAs. The default is set be the na.action setting of options, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.
start	starting values for the parameters in the linear predictor.
offset	this can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be NULL or a *non-positive* numeric vector of length equal to the number of cases. One or more offset terms can be included in the formula instead or as well, and if more than one is specified their sum is used. See model.offset.
control	a list of parameters for controlling the fitting process, passed to logbin.control.
model	a logical value indicating whether the *model frame* should be included as a component of the returned value.
warn	a logical indicating whether or not warnings should be provided for non-convergence or boundary values.
...	arguments to be used to form the default control argument if it is not supplied directly.

Details

logbin fits a generalised linear model (GLM) with a binomial error distribution and log link function. Predictors are assumed to be continuous, unless they are of class factor, or are character or logical (in which case they are converted to factors). Specifying a predictor as monotonic using the mono argument means that for continuous terms, the associated coefficient will be restricted to be non-negative, and for categorical terms, the coefficients will be non-decreasing in the order of the factor levels. This allows semi-parametric monotonic regression functions, in the form of unsmoothed step-functions. For smooth regression functions see logbin.smooth.

As well as allowing monotonicity constraints, the function is useful when a standard GLM routine, such as glm, fails to converge with a log-link binomial model. If glm does achieve successful convergence, and logbin converges to an interior point, then the two results will be identical. However, as illustrated in one of the examples below, glm may still experience convergence problems even when logbin converges to an interior point. Note that if logbin converges to a boundary point, then it may differ slightly from glm even if glm successfully converges, because of differences in the definition of the parameter space. logbin produces valid fitted values for covariate values within the Cartesian product of the observed range of covariate values, whereas glm produces valid fitted values just for the observed covariate combinations (assuming it successfully converges). This issue is only relevant when logbin converges to a boundary point.

The computational method is a combinatorial EM algorithm (Marschner, 2014) which accommodates the parameter contraints in the model and is more stable than

iteratively reweighted least squares. A collection of restricted parameter spaces is defined which covers the full parameter space, and the EM algorithm is applied within each restricted parameter space in order to find a collection of restricted maxima of the log-likelihood function, from which can be obtained the global maximum over the full parameter space. See Marschner and Gillett (2012) for further details.

Value

`logbin` returns an object of class `"logbin"`, which inherits from classes `"glm"` and `"lm"`. The function `summary.logbin` can be used to obtain or print a summary of the results.

The generic accessor functions `coefficients`, `fitted.values` and `residuals` can be used to extract various useful features of the value returned by `logbin`. Note that `effects` will not work.

An object of class `"logbin"` is a list containing the same components as an object of class `"glm"` (see the "Value" section of `glm`), but without `contrasts`, `qr`, `R` or `effects` components. It also includes:

`loglik` the maximised log-likelihood.

`aic.c` a small-sample corrected version of Akaike's *An Information Criterion* (Hurvich, Simonoff and Tsai, 1998). This is used by `logbin.smooth` to choose the optimal number of knots for smooth terms.

`xminmax` the minimum and maximum observed values for each of the continuous covariates, to help define the covariate space of the model.

As well as:

`np.coefficients` estimated coefficients associated with the non-positive parameterisation corresponding to the MLE.

`nn.x` non-negative model matrix associated with `np.coefficients`.

Note

Due to the way the covariate space is defined in the CEM algorithm, specifying interactions in the formula is not currently supported by `logbin`. 2-way interactions between factors can be included by calculating a new factor term that has levels corresponding to all possible combinations of the factor levels. See the Example.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

References

Hurvich, C. M., J. S. Simonoff and C.-L. Tsai (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(2): 271–293.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* 13(1): 179–192.

Examples

```
require(glm2)
data(heart)

#=====================================================
#  Model with periodic non-convergence when glm is used
#=====================================================

start.p <- sum(heart$Deaths) / sum(heart$Patients)

fit.glm <- glm(cbind(Deaths, Patients-Deaths) ~ factor(AgeGroup) +
  factor(Severity) + factor(Delay) + factor(Region),
  family = binomial(log), start = c(log(start.p), -rep(1e-4, 8)),
  data = heart, trace = TRUE, maxit = 100)

fit.logbin <- logbin(formula(fit.glm), data = heart, trace = 1)
## (Note that convergence may be sped up by specifying mono = c(1,2))

summary(fit.logbin)

#============================
#  Model with interaction term
#============================

heart$AgeSev <- 10 * heart$AgeGroup + heart$Severity

fit.logbin.int <- logbin(cbind(Deaths, Patients-Deaths) ~ factor(AgeSev) +
  factor(Delay) + factor(Region), data = heart, trace = 1, maxit = 100000)

summary(fit.logbin.int)
vcov(fit.logbin.int)
confint(fit.logbin.int)
summary(predict(fit.logbin.int, type = "response"))

anova(fit.logbin, fit.logbin.int, test = "Chisq")
```

`logbin.allref` *Parameterisation for CEM Algorithm*

Description

A workhorse function for `logbin`, `logbin.allref` takes the formula and data for a log-link binomial GLM and produces a list of all parameterisations needed for the associated CEM algorithm.

Usage

```
logbin.allref(object, data = environment(object), mono, start = NULL)
```

Arguments

`object`	a model `formula` or a `terms` object for the `logbin` model.
`data`	a data frame created with `model.frame`. If another sort of object, `model.frame` is called first.
`mono`	a vector indicating which terms should be restricted to have a monotonically non-decreasing relationship with the outcome.
`start`	starting values for the parameters in the linear predictor.

Details

In the CEM algorithms employed by `logbin`, the parameter space is partitioned into a collection of restricted parameter spaces (see Marschner, 2014). `logbin.allref` finds the list of possible parameterisations of each term in the model.

If a term `x` has a `TRUE` value for `is.factor(x)`, `is.character(x)` or `is.logical(x)`, it is considered to be a categorical covariate. This has a parameterisation for each level of the factor.

Otherwise the covariate is considered to be continuous, in which case it has two possible parameterisations, relating to the minimum and maximum observed values.

If a covariate is restricted to be monotonic via the `mono` argument, it has only one parameterisation.

`logbin` considers all possible combinations of the parameterisations of each covariate, and for each calls `logbin.design` to create the appropriate non-negative design matrix to be used in the EM algorithm.

Value

A list with components:

`allref`	a named list, with one component for each term in the model. Each component is itself a list, whose components are each of the parameterisations for that term.

	If `start` was specified, the first component for each term will correspond to the parameterisation specified by `start`.
`terms`	the `terms` component of `object`.
`data`	the object passed into the `data` argument, or the result of calling `model.frame` with `data`.
`monotonic`	a named logical vector indicating which components of `terms` are restricted to be monotonically non-decreasing.
`start.new`	a reparameterised version of `start`, corresponding to the first parameterisation in `allref`. `NULL` if `start` was not supplied.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

References

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* 13(1): 179–192.

See Also

`logbin`

`logbin.control` *Auxiliary for Controlling logbin Fitting*

Description

Auxiliary function for `logbin` fitting. Typically only used internally by `nplbin`, but may be used to construct a `control` argument to that function.

Usage

```
logbin.control(bound.tol = 1e-06, epsilon = 1e-08, maxit = 10000,
               trace = 0)
```

Arguments

`bound.tol`	positive tolerance specifying the interior of the parameter space. If the fitted model is more than `bound.tol` away from the boundary of the parameter space then it is assumed to be in the interior. This can allow the computational method to terminate early if an interior maximum is found. No early termination is attempted if `bound.tol = Inf`.

`epsilon` positive convergence tolerance ϵ; the estimates are considered to have converged when $\sqrt{\sum(\theta_{old} - \theta_{new})^2} / \sqrt{\sum \theta_{old}^2} < \epsilon$, where θ is the vector of parameter estimates. See `conv.test`.

`maxit` integer giving the maximum number of EM algorithm iterations for a given parameterisation.

`trace` number indicating level of output that should be produced. >= 1 gives output for each parameterisation, >= 2 gives output at each iteration.

Details

This is used similarly to `glm.control`. The `control` argument of `logbin` is by default passed to the `control` argument of `nplbin`.

When `trace` is greater than zero, calls to `cat` produce the output. Hence, `options( digits = *)` can be used to increase the precision.

Value

A list with components named as the arguments.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

See Also

`glm.control`, the equivalent function for `glm` fitting.

`nplbin`, the function used to fit `logbin` models.

Examples

```
## Variation on example(glm.control) :

evts <- c(18,17,15,20,10,20,25,13,12)
obs <- rep(30,9)
outcome <- gl(3,1,9)
treatment <- gl(3,3)
oo <- options(digits = 12)
logbin.D93X <- logbin(cbind(evts,obs-evts) ~ outcome + treatment,
  trace = 2, epsilon = 1e-2)
options(oo)
coef(logbin.D93X)
```

`logbin.design` *Construct Design Matrix for logbin Model*

Description

`logbin.design` constructs the design matrix for a `logbin` model, given a particular parameterisation.

This is a workhorse function — it would not normally be called directly.

Usage

```
logbin.design(terms, data, allref, design.ref)
```

Arguments

`terms`	`terms` component of object returned from a call to `logbin.allref` for the desired model.
`data`	`data` component of object returned from a call to `logbin.allref` for the desired model.
`allref`	`allref` component of object returned from a call to `logbin.allref` for the desired model.
`design.ref`	vector of indices for a particular parameterisation in `allref`. That is, each element corresponds to a term `x` in the model, and the value of the element indicates which item in the list `allref[[x]]` is the reference level in this parameterisation.

Details

In the CEM algorithm employed by `logbin`, we must consider the Cartesian product of all possible parameterisations. The list of these for each term in the model is constructed by a call to `logbin.allref`, and a list of all possible combinations created using `expand.grid`.

For a particular combination of reference levels, `logbin.design` constructs the associated design matrix by transforming `data`. Continuous covariates are transformed such that their minimum or maximum observed value corresponds to a transformed value of zero; categorical covariates are transformed by using either `contr.treatment` with a specified reference level or `contr.isotonic.rev` so that the levels are increasing in the specified order.

Value

A strictly non-negative design matrix to be passed to the relevant fitting function.

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

See Also

logbin.allref, model.matrix, contr.treatment, contr.isotonic.rev

---

logbin.smooth — *Smooth Log-Binomial Regression*

Description

logbin.smooth fits log-link binomial regression models using a stable CEM algorithm. It provides additional flexibility over logbin by allowing for smooth semi-parametric terms.

Usage

```
logbin.smooth(formula, mono = NULL, data, subset, na.action, offset,
              control = list(...), model = TRUE,
              model.logbin = FALSE, ...)
```

Arguments

formula	an object of class "formula" (or one that can be coerced into that class): a symbolic description of the model to be fitted. The details of model specification are given under "Details". The model must contain an intercept and at least one semi-parametric term, included by using the B or Iso functions. Note that 2nd-order terms (such as interactions) or above are not currently supported (see logbin).
mono	a vector indicating which terms in formula should be restricted to have a monotonically non-decreasing relationship with the outcome. May be specified as names or indices of the terms. Iso() terms are always monotonic.
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which logbin.smooth is called.
subset	an optional vector specifying a subset of observations to be used in the fitting process.
na.action	a function which indicates what should happen when the data contain NAs. The default is set be the na.action setting of options, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.

`offset`	this can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be `NULL` or a *non-positive* numeric vector of length equal to the number of cases. One or more `offset` terms can be included in the formula instead or as well, and if more than one is specified their sum is used. See `model.offset`.
`control`	a list of parameters for controlling the fitting process, passed to `logbin.control`.
`model`	a logical value indicating whether the *model frame* should be included as a component of the returned value.
`model.logbin`	a logical value indicating whether the fitted `logbin` object should be included as a component of the returned value.
`...`	arguments to be used to form the default `control` argument if it is not supplied directly.

Details

`logbin.smooth` performs the same fitting process as `logbin`, providing a stable maximum likelihood estimation procedure for log-link binomial GLMs, with the added flexibility of allowing semi-parametric `B` and `Iso` terms (note that `logbin.smooth` will stop with an error if no semi-parametric terms are specified in the right-hand side of the `formula`; `logbin` should be used instead).

The method partitions the parameter space associated with the semi-parametric part of the model into a sequence of constrained parameter spaces, and defines a fully parametric `logbin` model for each. The model with the highest log-likelihood is the MLE for the semi-parametric model (see Donoghoe and Marschner, 2015).

Value

An object of class `"logbin.smooth"`, which contains the same objects as class `"logbin"` (the same as `"glm"` objects, without `contrasts`, `qr`, `R` or `effects` components), as well as:

`model.logbin`	if `model.logbin` is `TRUE`; the `logbin` object for the fully parametric model corresponding to the fitted model.
`xminmax.smooth`	the minimum and maximum observed values for each of the smooth terms in the model, to help define the covariate space.
`full.formula`	the component from `interpret.logbin.smooth(formula)` that contains the `formula` term with any additional arguments to the `B` function removed.
`knots`	a named list containing the knot vectors for each of the smooth terms in the model.

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

References

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* 11(1): 91–108.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

See Also

logbin

Examples

```
## Simple example
x <- c(0.3, 0.2, 0.0, 0.1, 0.2, 0.1, 0.7, 0.2, 1.0, 0.9)
y <- c(5, 4, 6, 4, 7, 3, 6, 5, 9, 8)
m1 <- logbin.smooth(cbind(y, 10-y) ~ B(x, knot.range = 0:2), mono = 1,
  trace = 1)
m2 <- logbin.smooth(cbind(y, 10-y) ~ Iso(x))

plot(m1)
plot(m2)

summary(predict(m1, type = "response"))
summary(predict(m2, type = "response"))
```

---

logbin.smooth.allref — *Parameterisation for CEM Algorithm with Smooth Terms*

Description

A workhorse function for logbin.smooth, logbin.smooth.allref takes the formula and data for a log-link binomial GLM with smooth terms and produces a list of all parameterisations needed for the CEM algorithm associated with the semi-parametric part of the model.

Usage

```
logbin.smooth.allref(object, data = environment(object), mono,
                     logbin.smooth.spec, num.knots)
```

Arguments

object	terms object for the "fake.formula" associated with a logbin.smooth model (see interpret.logbin.smooth).
data	a data frame created with get_all_vars for the fake.formula.

`mono` a vector indicating which terms in `fake.formula` should be restricted to have a monotonically non-decreasing relationship with the outcome. May be specified as names or indices of the terms.

`logbin.smooth.spec` details of the smooth terms in the formula; this must be in the format returned by `interpret.logbin.smooth`.

`num.knots` a vector containing the number of interior knots to be used for each smooth term in the model (`NA` for `Iso` terms).

Details

Semi-parametric models in `logbin.smooth` use an extended CEM algorithm by partioning the parameter space associated with the smooth terms into a collection of restricted parameter spaces, each corresponding to a restricted fully parametric model that can be fit using `logbin`. This is a workhorse function that creates the list of possible parameterisations of each smooth term.

Isotonic terms and monotonic B-spline terms have only one parameterisation: where the maximum fitted value occurs at the maximum of the covariate range.

Unrestricted B-spline terms each have $k + 3$ parameterisations (where k is the number of internal knots), corresponding to the possible locations of the maximum of the smooth curve along the range of the covariate.

`logbin.smooth` considers all possible combinations of the number of knots for each smooth term, and all possible combinations of the associated parameterisations, and `logbin.smooth.design` creates the appropriate formula and design matrix to be used in the call to `logbin`.

Value

A list with components:

`allref` a named list, with one component for each smooth term in the model. Each component is itself a list, whose components are each of the parameterisations for that term.

`terms` the `terms` component of `object`.

`data` the object passed into the `data` argument.

`monotonic` a named logical vector indicating which components of `terms` are restricted to be monotonically non-decreasing.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

References

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* 11(1): 91–108.

See Also

logbin.smooth

`logbin.smooth.design` *Construct Design Matrix for logbin.smooth Model*

Description

`logbin.smooth.design` constructs the design matrix and associated formula for an `logbin.smooth` model, given a particular parameterisation, to be passed into `logbin` for fitting.

This is a workhorse function — it would not normally be called directly.

Usage

```
logbin.smooth.design(interpret, allref, design.knots, design.param)
```

Arguments

`interpret`	the object returned by running `interpret.logbin.smooth` for the desired model, containing details of the smooth components.
`allref`	the object returned by running `logbin.smooth.allref` for the desired model.
`design.knots`	a vector containing the number of internal knots for each smooth term (`NA` for `Iso` terms).
`design.param`	a vector of indices for a particular parameterisation in `allref$allref`. Each element corresponds to a smooth term in the model, and the value indicates which item in the associated list is the reference level for this parameterisation.

Details

For a particular combination of reference levels, `logbin.smooth.design` constructs the associated design matrix and formula. Specifically, for `Iso` smooth components, it creates the matrix of indicator covariates for increments between levels. For `B` smooth components, it creates the basis functions using `splineDesign` and removing the column associated with the reference level (see Donoghoe and Marschner, 2015).

The `formula` component is altered to include the terms in the design matrix, and `allref$monotonic` is altered such that all of the smooth coefficients are restricted to be non-negative, as required.

Value

A list with components:

`formula`	an updated version of `interpret$full.formula`, with smooth terms removed and replaced by the names of their associated basis components.
`data`	an updated version of `interpret$data`, with columns for the basis functions of the smooth terms added.
`monotonic`	an updated version of `allref$monotonic`, such that the coefficients associated with the smooth terms for this parameterisation are constrained to be non-negative.
`knots`	a list, with one component for each smooth term, containing the knot vector for that term (`NA` for `Iso` terms).

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

References

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* 11(1): 91–108.

See Also

`logbin.smooth`, `interpret.logbin.smooth`, `logbin.smooth.allref`, `logbin.design`

`nplbin` *Non-Positive Log-Binomial Regression*

Description

Finds the maximum likelihood estimate of a log-link binomial GLM using an EM algorithm, where each of the coefficients in the linear predictor is restricted to be non-positive.

Usage

```
nplbin(y, x, offset, start, control = list())
```

Arguments

`y`	binomial response. May be a single column of 0/1 or two columns, giving the number of successes and failures.
`x`	non-negative covariate matrix.
`offset`	non-positive additive offset vector. The default is a vector of zeros.
`start`	starting values for the parameter estimates. All elements must be less than or equal to `-control$bound.tol`.
`control`	a `logbin.control` object, which controls the fitting process.

Details

This is a workhorse function for `logbin`, and runs the EM algorithm to find the constrained non-positive MLE associated with a log-link binomial GLM. See Marschner and Gillett (2012) for full details.

Value

A list containing the following components

`coefficients`	the constrained non-positive maximum likelihood estimate of the parameters.
`residuals`	the residuals at the MLE, that is `y - fitted.values`
`fitted.values`	the fitted mean values.
`rank`	the number of parameters in the model (named "`rank`" for compatibility — we assume that models have full rank)
`family`	included for compatibility — will always be `binomial(log)`.
`linear.predictors`	the linear fit on link scale.
`deviance`	up to a constant, minus twice the maximised log-likelihood.
`aic`	a version of Akaike's *An Information Criterion*, minus twice the maximised log-likelihood plus twice the number of parameters.
`aic.c`	a small-sample corrected version of Akaike's *An Information Criterion* (Hurvich, Simonoff and Tsai, 1998).
`null.deviance`	the deviance for the null model, comparable with `deviance`. The null model will include the offset and an intercept.
`iter`	the number of iterations of the EM algorithm used.
`weights`	included for compatibility — a vector of ones.
`prior.weights`	the number of trials associated with each binomial response.
`df.residual`	the residual degrees of freedom.
`df.null`	the residual degrees of freedom for the null model.
`y`	the `y` vector used.

converged	logical. Did the EM algorithm converge (according to conv.test)?
boundary	logical. Is the MLE on the boundary of the parameter space — i.e. are any of the coefficients < control$bound.tol?
loglik	the maximised log-likelihood.
nn.design	the non-negative x matrix used.

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>.

This function is based on code from Marschner and Gillett (2012) written by Alexandra Gillett.

References

Hurvich, C. M., J. S. Simonoff and C.-L. Tsai (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(2): 271–293.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* 13(1): 179–192.

---

plot.logbin.smooth *Default logbin.smooth Plotting*

Description

Takes a fitted logbin.smooth object produced by logbin.smooth and plots the component smooth functions that make it up, for specified values of the other covariates.

Usage

```
## S3 method for class 'logbin.smooth'
plot(x, type = c("response", "link"), at = data.frame(),
     knotlines = TRUE, nobs = 1000, ...)
```

Arguments

x	a fitted logbin.smooth object as produced by logbin.smooth.
type	the type of prediction required. Note that, unlike predict.logbin.smooth, "terms" is not a valid option.
at	a data frame containing the values at which the prediction should be evaluated. The columns must contain the covariates in the model, and several rows may be provided (in which case, multiple lines are drawn on the same plot). Cannot be missing or NULL.

`knotlines`	logical; if vertical lines should be drawn on the plot to indicate the locations of the knots for B-spline terms.
`nobs`	the number of points which should be used to create the curve. These are placed evenly along the range of the observed covariate values from the original model.
`...`	other graphics parameters to pass on to plotting commands (note: some will not work).

Details

For each smooth covariate in the model of `x`, `predict.logbin.smooth` is used to obtain predicted values for the range of that covariate, with the other covariates remaining fixed at their values given in `at`. Several rows may be provided in `at`, in which case, one curve is drawn for each, and they are coloured using `rainbow(nrow(at))`. If the model contains a single smooth covariate and no other covariates, `at` may be provided as an empty data frame, `data.frame()`.

Value

The function simply generates plots.

Note

If this function is too restrictive, it may be easier to use `predict.logbin.smooth` to get predictions for the dataset of your choice, and do the plotting manually.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

See Also

`logbin.smooth`, `predict.logbin.smooth`

Examples

```
## For an example, see example(logbin.smooth)
```

`predict.logbin` *Predict Method for logbin Fits*

Description

Obtains predictions from a fitted `logbin` object.

Usage

```
## S3 method for class 'logbin'
predict(object, newdata = NULL, type = c("link", "response",
        "terms"), terms = NULL, na.action = na.pass,
        checkminmax = TRUE, ...)
```

Arguments

`object`	a fitted object of class inheriting from `"logbin"`.
`newdata`	optionally, a data frame in which to look for variables with which to predict. If omitted, the fitted linear predictors are used.
`type`	the type of prediction required. The default is on the scale of the linear predictors; the alternative `"response"` is on the scale of the response variable. The `"terms"` option returns a matrix giving the fitted values of each term in the model formula on the linear predictor scale. The value of this argument can be abbreviated.
`terms`	with `type = "terms"` by default all terms are returned. A character vector specifies which terms are to be returned.
`na.action`	function determining what should be done with missing values in `newdata`. The default is to predict `NA`.
`checkminmax`	logical indicating whether or not values of continuous covariates in `newdata` should be checked to ensure they lie within the covariate space associated with the fitted model. Otherwise predicted values could lie outside the parameter space.
`...`	further arguments passed to or from other methods.

Details

If `newdata` is omitted the predictions are based on the data used for the fit. In that case how cases with missing values in the original fit are treated is determined by the `na.action` argument of that fit. If `na.action = na.omit`, omitted cases will not appear in the residuals. If `na.action = na.exclude` they will appear, with residual value `NA`. See also `napredict`.

Value

A vector or matrix of predictions. For `type = "terms"`, this is a matrix with a column per term, and may have an attribute `"constant"`.

Note

Variables are first looked for in `newdata` and then searched for in the usual way (which will include the environment of the formula used in the fit). A warning will be given if the variables found are not of the same length as those in `newdata` if it was supplied.

Author(s)

Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

See Also

logbin

predict.glm for the equivalent method for models fit using glm.

Examples

```
## For an example, see example(logbin)
```

---

predict.logbin.smooth *Predict Method for logbin.smooth Fits*

Description

Obtains predictions from a fitted logbin.smooth object.

Usage

```
## S3 method for class 'logbin.smooth'
predict(object, newdata = NULL, type = c("link", "response",
        "terms"), terms = NULL, na.action = na.pass, ...)
```

Arguments

object	a fitted object of class inheriting from "logbin.smooth".
newdata	optionally, a data frame in which to look for variables with which to predict. If omitted, the fitted linear predictors are used.
type	the type of prediction required. The default is on the scale of the linear predictors; the alternative "response" is on the scale of the response variable. The "terms" option returns a matrix giving the fitted values of each term in the model formula on the linear predictor scale. The value of this argument can be abbreviated.
terms	with type = "terms" by default all terms are returned. A character vector specifies which terms are to be returned.
na.action	function determining what should be done with missing values in newdata. The default is to predict NA.
...	further arguments passed to or from other methods.

Details

`predict.logbin.smooth` constructs the underlying basis functions for smooth variables in `newdata` and runs `predict.logbin` to obtain predictions. Note that if values of smooth covariates in `newdata` are outside the covariate space of `object`, an error will be returned.

If `newdata` is omitted, the predictions are based on the data used for the fit. In that case how cases with missing values in the original fit are treated is determined by the `na.action` argument of that fit. If `na.action = na.omit`, omitted cases will not appear in the residuals, whereas if `na.action = na.exclude` they will appear, with residual value `NA`. See also `napredict`.

Value

A vector or matrix of predictions. For `type = "terms"`, this is a matrix with a column per term, and may have an attribute `"constant"`.

Note

Variables are first looked for in `newdata` and then searched for in the usual way (which will include the environment of the formula used in the fit). A warning will be given if the variables found are not of the same length as those in `newdata` if it was supplied.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

See Also

`logbin.smooth`, `predict.logbin`

`predict.glm` for the equivalent method for models fit using `glm`.

Examples

```
## For an example, see example(logbin.smooth)
```

`summary.logbin` *Summarising logbin Model Fits*

Description

These functions are all `methods` for class `logbin` or `summary.logbin` objects.

Usage

```
## S3 method for class 'logbin'
summary(object, correlation = FALSE, ...)

## S3 method for class 'summary.logbin'
print(x, digits = max(3L, getOption("digits") - 3L),
      signif.stars = getOption("show.signif.stars"), ...)
```

Arguments

`object`	an object of class `"logbin"`, usually from a call to `logbin` or `logbin.smooth`.
`x`	an object of class `"summary.logbin"`, usually from a call to `summary.logbin`.
`correlation`	logical; if `TRUE`, the correlation matrix of the estimated parameters is returned and printed.
`digits`	the number of significant digits to use when printing.
`signif.stars`	logical; if `TRUE`, 'significance stars' are printed for each coefficient.
`...`	further arguments passed to or from other methods.

Details

These perform the same function as `summary.glm` and `print.summary.glm`, producing similar results for `logbin` models. `print.summary.logbin` additionally prints the small-sample corrected AIC (`aic.c`), and the number of EM iterations for the parameterisation corresponding to the MLE.

The dispersion used in calculating standard errors is fixed as `1`.

Value

`summary.logbin` returns an object of class `"summary.logbin"`, a list with components

`call`	the component from `object`.
`family`	the component from `object`.
`deviance`	the component from `object`.
`aic`	the component from `object`.
`aic.c`	the component from `object`.
`df.residual`	the component from `object`.
`null.deviance`	the component from `object`.
`df.null`	the component from `object`.
`iter`	the component from `object`.
`deviance.resid`	the deviance residuals: see `residuals.glm`.

`coefficients`	the matrix of coefficients, standard errors, z-values and p-values.
`aliased`	included for compatibility — always `FALSE`.
`dispersion`	the inferred/estimated dispersion.
`df`	included for compatibility — a 3-vector of the number of coefficients, the number of residual degrees of freedom, and the number of coefficients (again).
`cov.unscaled`	the unscaled (`dispersion = 1`) estimated covariance matrix of the estimated coefficients. `NaN` if `object$boundary == TRUE`.
`cov.scaled`	ditto, scaled by `dispersion`.
`correlation`	if `correlation` is `TRUE`, the estimated correlations of the estimated coefficients. `NaN` if `object$boundary == TRUE`.

Note

If `object$boundary == TRUE`, the standard errors of the coefficients are not valid, and a matrix of `NaN`s is returned by `vcov.logbin`.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

See Also

`logbin`, `summary.glm`

Examples

```
## For examples see example(logbin)
```

`vcov.logbin` *Calculate Variance-Covariance Matrix for a Fitted logbin Model Object*

Description

Returns the variance-covariance matrix of the main parameters of a fitted `logbin` model object.

Usage

```
## S3 method for class 'logbin'
vcov(object, ...)
```

Arguments

`object`	an object of class `"logbin"`, usually from a call to `logbin` or `logbin.smooth`.
`...`	additional arguments for method functions.

Details

An equivalent method to `vcov`, to use with `logbin` models.

Value

A matrix of the estimated covariances between the parameter estimates in the linear or non-linear predictor of the model. This should have row and column names corresponding to the parameter names given by the `coef` method.

Note

If `object$boundary == TRUE`, the standard errors of the coefficients are not valid, and a matrix of `NaN`s is returned.

Author(s)

Mark W. Donoghoe <`mark.donoghoe@mq.edu.au`>

See Also

`summary.logbin`, `vcov.glm`

Examples

```
## For an example see example(logbin)
```

B

`addreg` package documentation

This appendix contains the documentation for the R package `addreg`, which implements methods for fitting semi-parametric additive Poisson, binomial and negative binomial regression models as described in Marschner (2010), and Chapters 3, 4 and 5 of this thesis.

The package is available online from the Comprehensive R Archive Network:

Donoghoe, M. W. (2015a). *addreg: Additive Regression for Discrete Data*. R package version 2.0. URL: http://CRAN.R-project.org/package=addreg.

The function `nnpois`, which implements the constrained additive Poisson regression EM algorithm, is based on code written by Alexandra Gillett and published as supplementary material to Marschner, Gillett, and O'Connell (2012). All other functions were written entirely by the candidate.

Package 'addreg'

Title Additive Regression for Discrete Data

Description Methods for fitting identity-link GLMs and GAMs to discrete data, using EM-type algorithms with more stable convergence properties than standard methods.

Version 2.0

Date 2015-05-12

Author Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

Maintainer Mark W. Donoghoe <mark.donoghoe@mq.edu.au>

Depends R ($\geq$ 3.0.1)

Imports splines, combinat, glm2

License GPL ($\geq$ 2)

LazyData true

`addreg-package` *Additive Regression for Discrete Data*

Description

Methods for fitting identity-link GLMs and GAMs to discrete data, using EM-type algorithms with more stable convergence properties than standard methods.

Details

Package:	addreg
Type:	Package
Version:	2.0
Date:	2015-05-12
License:	GPL ($\geq$ 2)

This package provides methods to fit generalised linear models (GLMs) and generalised additive models (GAMs) with identity link functions to discrete data using binomial, Poisson and negative binomial models. It is planned that future versions will incorporate other types of discrete data models, such as multinomial regression.

The package has two primary functions: `addreg` and `addreg.smooth`, together with various supporting functions. It is useful in two main situations. The first is when a standard GLM routine, such as `glm`, fails to converge with such a model. The second is when a flexible semi-parametric component is desired in these models. One of the main purposes of this package is to provide parametric and semi-parametric adjustment of risk differences and rate differences.

The computational method is a combinatorial EM algorithm (Marschner, 2014), which accommodates the parameter constraints and is more stable than iteratively reweighted least squares. A collection of restricted parameter spaces is defined which covers the full parameter space, and the EM algorithm is applied within each restricted parameter space in order to find a collection of restricted maxima of the log-likelihood function, from which can be obtained the global maximum over the full parameter space.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

Maintainer: Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

References

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* 80: 184–196.

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* 11(1): 91–108.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* 19(3): 666–683.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

See Also

`glm`

Examples

```
## For examples, see example(addreg) and example(addreg.smooth)
```

`addbin` *Fitting Additive Binomial Regression Models*

Description

Workhorse function for `addreg` with `binomial` family.

Usage

```
addbin(y, x, start = NULL, control = list(), allref)
```

Arguments

`y`	binomial response. May be a single column of 0/1 or two columns, giving the number of successes and failures.
`x`	non-negative design matrix. Must have an intercept column.
`start`	starting values for the parameters in the linear predictor.
`control`	list of parameters for controlling the fitting process, passed to `addreg.control`.
`allref`	a list of all parameterisations for this model, obtained from `addreg.allref`.

Details

An additive binomial fit can be converted into an additive Poisson fit via the multinomial–Poisson transformation (Baker, 1994). This function transforms the data as described by Donoghoe and Marschner (2014) and passes it to `addreg` with a Poisson family to get the maximum likelihood estimate. The coefficients (and other values) from the Poisson model are transformed back to relate to the additive binomial model.

This is a workhorse function for `addreg` when a `binomial` family is specified. It would not usually be called directly.

Value

A list of (most of) the components needed for an object of class `"addreg"`; see `addreg` for details.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

References

Baker, S. G. (1994). The multinomial–Poisson transformation. *The Statistician* 43(4): 495–504.

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* 80: 184–196.

See Also

`addreg`

`addreg` *Additive Regression for Discrete Data*

Description

`addreg` fits additive (identity-link) Poisson, negative binomial and binomial regression models using a stable combinatorial EM algorithm.

Usage

```
addreg(formula, mono = NULL, family, data, standard, subset,
       na.action, start = NULL, offset, control = list(...),
       model = TRUE, warn = TRUE, ...)
```

Arguments

`formula`	an object of class `"formula"` (or one that can be coerced into that class): a symbolic description of the model to be fitted. The details of model specification are given under "Details". Note that the model must contain an intercept, and 2nd-order terms (such as interactions) or above are currently not supported — see "Note".
`mono`	a vector indicating which terms in `formula` should be restricted to have a monotonically non-decreasing relationship with the outcome. May be specified as names or indices of the terms.
`family`	a description of the error distribution to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function (see `family` for details of family functions), but here it is restricted to be `poisson`, `negbin1` or `binomial` family with `identity` link.
`data`	an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`, typically the environment from which `addreg` is called.
`standard`	a numeric vector of length equal to the number of cases, where each element is a positive constant that (multiplicatively) standardises the fitted value of the corresponding element of the response vector. Ignored for binomial family (two-column specification of response should be used instead).
`subset`	an optional vector specifying a subset of observations to be used in the fitting process.
`na.action`	a function which indicates what should happen when the data contain `NA`s. The default is set be the `na.action` setting of `options`, and is `na.fail` if that is unset. The 'factory-fresh' default is `na.omit`. Another possible value is `NULL`, no action. Value `na.exclude` can be useful.
`start`	starting values for the parameters in the linear predictor, also with the starting value for the `scale` as the last element when `family = negbin1`.
`offset`	this can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be `NULL` or a *non-negative* numeric vector of length equal to the number of cases. One or more `offset` terms can be included in the formula instead or as well, and if more than one is specified their sum is used. See `model.offset`. Ignored for binomial family.
`control`	list of parameters for controlling the fitting process, passed to `addreg.control`.

`model`	a logical value indicating whether the *model frame* (and, for binomial models, the equivalent Poisson model) should be included as a component of the returned value.
`warn`	a logical indicating whether or not warnings should be provided for non-convergence or boundary values.
`...`	arguments to be used to form the default `control` argument if it is not supplied directly.

Details

`addreg` fits a generalised linear model (GLM) with a Poisson or binomial error distribution and identity link function, as well as additive NegBin I models (which are not GLMs). Predictors are assumed to be continuous, unless they are of class `factor`, or are character or logical (in which case they are converted to `factor`s). Specifying a predictor as monotonic using the `mono` argument means that for continuous terms, the associated coefficient will be restricted to be non-negative, and for categorical terms, the coefficients will be non-decreasing in the order of the factor `levels`. This allows semi-parametric monotonic regression functions, in the form of unsmoothed step-functions. For smooth regression functions see `addreg.smooth`.

As well as allowing monotonicity constraints, the function is useful when a standard GLM routine, such as `glm`, fails to converge with an identity-link Poisson or binomial model. If `glm` does achieve successful convergence, and `addreg` converges to an interior point, then the two results will be identical. However, `glm` may still experience convergence problems even when `addreg` converges to an interior point. Note that if `addreg` converges to a boundary point, then it may differ slightly from `glm` even if `glm` successfully converges, because of differences in the definition of the parameter space. `addreg` produces valid fitted values for covariate values within the Cartesian product of the observed range of covariate values, whereas `glm` produces valid fitted values just for the observed covariate combinations (assuming it successfully converges). This issue is only relevant when `addreg` converges to a boundary point.

The computational method is a combinatorial EM algorithm (Marschner, 2014), which accommodates the parameter contraints in the model and is more stable than iteratively reweighted least squares. A collection of restricted parameter spaces is defined which covers the full parameter space, and the EM algorithm is applied within each restricted parameter space in order to find a collection of restricted maxima of the log-likelihood function, from which can be obtained the global maximum over the full parameter space. See Marschner (2010) and Donoghoe and Marschner (2014) for further details.

Value

`addreg` returns an object of class `"addreg"`, which inherits from classes `"glm"` and `"lm"`. The function `summary.addreg` can be used to obtain or print a summary of the results.

The generic accessor functions `coefficients`, `fitted.values` and `residuals` can be used to extract various useful features of the value returned by `addreg`. Note

that `effects` will not work.

An object of class `"addreg"` is a list containing the same components as an object of class `"glm"` (see the "Value" section of `glm`), but without `contrasts`, `qr`, `R` or `effects` components. It also includes:

`loglik`	the maximised log-likelihood.
`aic.c`	a small-sample corrected version of Akaike's *An Information Criterion* (Hurvich, Simonoff and Tsai, 1998). This is used by `addreg.smooth` to choose the optimal number of knots for smooth terms.
`xminmax`	the minimum and maximum observed values for each of the continuous covariates, to help define the covariate space of the model.

As well as, for Poisson and negative binomial models:

`nn.coefficients`	estimated coefficients associated with the non-negative parameterisation corresponding to the MLE.
`nn.x`	non-negative model matrix associated with `nn.coefficients`.
`standard`	the `standard` argument.

Or, for binomial models:

`model.addpois`	if requested, the `addreg` object for the associated identity-link Poisson model.

The `scale` component of the result is fixed at 1 for Poisson and binomial models, and is the constant overdispersion parameter for negative binomial models (that is, `scale` $= 1 + \phi$) where $Var(\mu) = (1 + \phi)\mu$).

Note

Due to the way the covariate space is defined in the CEM algorithm, specifying interactions in the formula is not currently supported by `addreg`. 2-way interactions between factors can be included by calculating a new factor term that has levels corresponding to all possible combinations of the factor levels. See the Example.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

References

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* 80: 184–196.

Hurvich, C. M., J. S. Simonoff and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion.

Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60(2): 271–293.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* 19(3): 666–683.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

Examples

```
require(glm2)
data(crabs)

#============================================================
#  Poisson model with periodic non-convergence when glm is used
#============================================================

crabs.boot <- crabs[crabs$Rep1,-c(5:6)]
crabs.boot$width.shifted <- crabs.boot$Width - min(crabs$Width)

fit.glm <- glm(Satellites ~ width.shifted + factor(Dark) +
  factor(GoodSpine), family = poisson(identity), data = crabs.boot,
  start = rep(1,4), control = glm.control(trace = TRUE))

fit.addreg <- addreg(formula(fit.glm), family = poisson,
  data = crabs.boot, trace = 1)
summary(fit.addreg)
vcov(fit.addreg)
confint(fit.addreg)
summary(predict(fit.addreg), type = "response")

fit.addreg2 <- addreg(update(formula(fit.glm), ~ . - factor(GoodSpine)),
  family = poisson, data = crabs.boot, trace = 1)
anova(fit.addreg2, fit.addreg, test = "LRT")

# Account for overdispersion (use start to speed it up a little)
fit.addreg.od <- addreg(Satellites ~ factor(Dark) + factor(GoodSpine),
  family = negbin1, data = crabs.boot, trace = 1,
  start = c(4.3423675,-2.4059273,-0.4531984,5.969648))
summary(fit.addreg.od)
```

`addreg.allref` *Parameterisation for CEM Algorithm*

Description

A workhorse function for `addreg`, `addreg.allref` takes the formula and data for an identity-link GLM and produces a list of all parameterisations needed for the associated CEM algorithm.

Usage

```
addreg.allref(object, data = environment(object), mono, family,
              start = NULL)
```

Arguments

object	a model formula or a terms object for the addreg model.
data	a data frame created with model.frame. If another sort of object, model.frame is called first.
mono	a vector indicating which terms should be restricted to have a monotonically non-decreasing relationship with the outcome.
family	the result of a call to a family function. Its component $family must be one of "poisson", "negbin1" or "binomial".
start	starting values for the parameters in the linear predictor.

Details

In the CEM algorithms employed by addreg, the parameter space is partitioned into a collection of restricted parameter spaces (see Marschner, 2014). addreg.allref finds the list of possible parameterisations of each term in the model.

If a term x has a TRUE value for is.factor(x), is.character(x) or is.logical(x), it is considered to be a categorical covariate. For Poisson and negative binomial models, this has a parameterisation for each level of the factor, and for binomial models, every permutation of the levels must be considered (see Donoghoe and Marschner, 2014).

Otherwise the covariate is considered to be continuous, in which case it has two possible parameterisations, relating to the minimum and maximum observed values.

If a covariate is restricted to be monotonic via the mono argument, it has only one parameterisation.

The addreg function considers all possible combinations of the parameterisations of each covariate, and uses addreg.design to create the appropriate non-negative design matrix to be used in the EM algorithm.

Value

A list with components:

allref	a named list, with one component for each term in the model. Each component is itself a list, whose components are each of the parameterisations for that term. If start was specified, the first component for each term will correspond to the parameterisation specified by start.
terms	the terms component of object.
data	the object passed into the data argument, or the result of calling model.frame with data.

`monotonic` a named logical vector indicating which components of `terms` are restricted to be monotonically non-decreasing.

`start.new` a reparameterised version of `start`, corresponding to the first parameterisation in `allref`. `NULL` if `start` was not supplied.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

References

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* 80: 184–196.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

See Also

`addreg`

`addreg.control` *Auxiliary for Controlling addreg Fitting*

Description

Auxiliary function for `addreg` fitting. Typically only used internally by `nnpois`, `nnnegbin` and `addbin`, but may be used to construct a `control` argument to these functions.

Usage

```
addreg.control(bound.tol = 1e-06, epsilon = 1e-10, maxit = 10000,
               trace = 0)
```

Arguments

`bound.tol` positive tolerance specifying the interior of the parameter space. If the fitted model is more than `bound.tol` away from the boundary of the parameter space then it is assumed to be in the interior. This can allow the computational method to terminate early if an interior maximum is found. No early termination is attempted if `bound.tol = Inf`.

`epsilon` positive convergence tolerance ϵ; the estimates are considered to have converged when $\sqrt{\sum(\theta_{old} - \theta_{new})^2} / \sqrt{\sum \theta_{old}^2} < \epsilon$, where θ is the vector of parameter estimates. See `conv.test`.

`maxit`	integer giving the maximum number of EM algorithm iterations for a given parameterisation.
`trace`	number indicating level of output that should be produced. >= 1 gives output for each parameterisation, >= 2 gives output at each iteration.

Details

This is used similarly to `glm.control`. The `control` argument of `addreg` is by default passed to the `control` argument of `nnpois`, `nnnegbin` or `addbin`.

When `trace` is greater than zero, calls to `cat` produce the output. Hence, `options(digits = *)` can be used to increase the precision.

Value

A list with components named as the arguments.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`glm.control`, the equivalent function for `glm` fitting.

`nnpois`, `nnnegbin` and `addbin`, the functions used to fit `addreg` models.

Examples

```
## Variation on example(glm.control) :

counts <- c(18,17,15,20,10,20,25,13,12)
outcome <- gl(3,1,9)
treatment <- gl(3,3)
oo <- options(digits = 12)
addreg.D93X <- addreg(counts ~ outcome + treatment, family = poisson,
  trace = 2, epsilon = 1e-2)
options(oo)
coef(addreg.D93X)
```

`addreg.design` *Construct Design Matrix for addreg Model*

Description

`addreg.design` constructs the design matrix for an `addreg` model, given a particular parameterisation.

This is a workhorse function — it would not normally be called directly.

Usage

```
addreg.design(terms, data, allref, design.ref)
```

Arguments

`terms`	`terms` component of object returned from a call to `addreg.allref` for the desired model.
`data`	`data` component of object returned from a call to `addreg.allref` for the desired model.
`allref`	`allref` component of object returned from a call to `addreg.allref` for the desired model.
`design.ref`	vector of indices for a particular parameterisation in `allref`. That is, each element corresponds to a term `x` in the model, and the value of the element indicates which item in the list `allref[[x]]` is the reference level in this parameterisation.

Details

In the CEM algorithm employed by `addreg`, we must consider the Cartesian product of all possible parameterisations. The list of these for each term in the model is constructed by a call to `addreg.allref`, and a list of all possible combinations created using `expand.grid`.

For a particular combination of reference levels, `addreg.design` constructs the associated design matrix by transforming `data`. Continuous covariates are transformed such that their minimum or maximum observed value corresponds to a transformed value of zero; categorical covariates are transformed by using either `contr.treatment` with a specified reference level or `contr.isotonic` so that the levels are increasing in the specified order.

Value

A strictly non-negative design matrix to be passed to the relevant fitting function.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

See Also

`addreg.allref`, `model.matrix`, `contr.treatment`, `contr.isotonic`

`addreg.smooth` *Smooth Additive Regression for Discrete Data*

Description

`addreg.smooth` fits additive (identity-link) Poisson, negative binomial and binomial regression models using a stable EM algorithm. It provides additional flexibility over `addreg` by allowing for semi-parametric terms.

Usage

```
addreg.smooth(formula, mono = NULL, family, data, standard, subset,
              na.action, offset, control = list(...), model = TRUE,
              model.addreg = FALSE, ...)
```

Arguments

`formula`	an object of class `"formula"` (or one that can be coerced into that class): a symbolic description of the model to be fitted. The details of model specification are given under "Details". The model must contain an intercept and at least one semi-parametric term, included by using the `B` or `Iso` functions. Note that 2nd-order terms (such as interactions) or above are not currently supported (see `addreg`).
`mono`	a vector indicating which terms in `formula` should be restricted to have a monotonically non-decreasing relationship with the outcome. May be specified as names or indices of the terms. `Iso()` terms are always monotonic.
`family`	a description of the error distribution to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function (see `family` for details of family functions), but here it is restricted to be `poisson`, `negbin1` or `binomial` family with `identity` link.
`data`	an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`, typically the environment from which `addreg.smooth` is called.
`standard`	a numeric vector of length equal to the number of cases, where each element is a positive constant that (multiplicatively) standardises the fitted value of the corresponding element of the response vector. Ignored for binomial family (the two-column specification of response should be used instead).
`subset`	an optional vector specifying a subset of observations to be used in the fitting process.

na.action	a function which indicates what should happen when the data contain NAs. The default is set be the na.action setting of options, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.
offset	this can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be NULL or a *non-negative* numeric vector of length equal to the number of cases. One or more offset terms can be included in the formula instead or as well, and if more than one is specified their sum is used. See model.offset. Ignored for binomial family.
control	list of parameters for controlling the fitting process, passed to addreg.control.
model	a logical value indicating whether the *model frame* (and, for binomial models, the equivalent Poisson model) should be included as a component of the returned value.
model.addreg	a logical value indicating whether the fitted addreg object should be included as a component of the returned value.
...	arguments to be used to form the default control argument if it is not supplied directly.

Details

addreg.smooth performs the same fitting process as addreg, providing a stable maximum likelihood estimation procedure for identity-link Poisson, negative binomial or binomial models, with the added flexibility of allowing semi-parametric B and Iso terms (note that addreg.smooth will stop with an error if no semi-parametric terms are specified in the right-hand side of the formula; addreg should be used instead).

The method partitions the parameter space associated with the semi-parametric part of the model into a sequence of constrained parameter spaces, and defines a fully parametric addreg model for each. The model with the highest log-likelihood is the MLE for the semi-parametric model (see Donoghoe and Marschner, 2015).

Value

An object of class "addreg.smooth", which contains the same objects as class "addreg" (the same as "glm" objects, without contrasts, qr, R or effects components), as well as:

model.addreg	if model.addreg is TRUE; the addreg object for the fully parametric model corresponding to the fitted model.
xminmax.smooth	the minimum and maximum observed values for each of the smooth terms in the model, to help define the covariate space.

`full.formula` the component from `interpret.addreg.smooth(formula)` that contains the `formula` term with any additional arguments to the `B` function removed.

`knots` a named list containing the knot vectors for each of the smooth terms in the model.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

References

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* 11(1): 91–108.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* 24(6): 921–940.

See Also

`addreg`

Examples

```
## Simple example
dat <- data.frame(
  x1 = c(3.2,3.3,3.4,7.9,3.8,0.7,2.0,5.4,8.4,3.0,1.8,5.6,5.5,9.0,8.2),
  x2 = c(1,0,0,1,0,1,0,0,0,0,1,0,1,1,0),
  n = c(6,7,5,9,10,7,9,6,6,7,7,8,6,8,10),
  y = c(2,1,2,6,3,1,2,2,4,4,1,2,5,7,7))
m1 <- addreg.smooth(cbind(y, n-y) ~ B(x1, knot.range = 1:3) + factor(x2),
  mono = 1, data = dat, family = binomial, trace = 1)

plot(m1, at = data.frame(x2 = 0:1))
points(dat$x1, dat$y / dat$n)
```

---

addreg.smooth.allref *Parameterisation for CEM Algorithm with Smooth Terms*

Description

A workhorse function for `addreg.smooth`, `addreg.smooth.allref` takes the formula and data for an identity-link GLM with smooth terms and produces a list of all parameterisations needed for the CEM algorithm associated with the semi-parametric part of the model.

Usage

```
addreg.smooth.allref(object, data = environment(object), mono,
                     family, addreg.smooth.spec, num.knots)
```

Arguments

`object`	`terms` object for the "`fake.formula`" associated with an `addreg.smooth` model (see `interpret.addreg.smooth`).
`data`	a data frame created with `get_all_vars` for the `fake.formula`.
`mono`	a vector indicating which terms in `fake.formula` should be restricted to have a monotonically non-decreasing relationship with the outcome. May be specified as names or indices of the terms.
`family`	the `family` object for the `addreg.smooth` model.
`addreg.smooth.spec`	details of the smooth terms in the formula; must be a list in the format returned by `interpret.addreg.smooth`.
`num.knots`	a vector containing the number of interior knots to be used for each smooth term in the model (`NA` for `Iso` terms).

Details

Semi-parametric models in `addreg.smooth` use an extended CEM algorithm by partioning the parameter space associated with the smooth terms into a collection of restricted parameter spaces, each corresponding to a restricted fully parametric model that can be fitted using `addreg`. The workhorse function `addreg.smooth.allref` creates the list of possible parameterisations of each smooth term.

Isotonic terms and monotonic B-spline terms have only one parameterisation: where the minimum fitted value occurs at the minimum of the covariate range.

For Poisson and negative binomial models, general B-spline terms have $k + 3$ parameterisations each (where k is the number of internal knots), corresponding to the possible locations of the minimum of the smooth curve along the range of the covariate.

For binomial models, general B-spline terms have $(k + 3)!$ parameterisations, corresponding to the permutations of the coefficients.

`addreg.smooth` considers all possible combinations of the number of knots for each smooth term, and all possible combinations of the associated parameterisations, and `addreg.smooth.design` creates the appropriate formula and design matrix to be used in the call to `addreg`.

Value

A list with components:

`allref`	a named list, with one component for each smooth term in the model. Each component is itself a list, whose components are each of the parameterisations for that term.

terms	the terms component of object.
data	the object passed into the data argument.
monotonic	a named logical vector indicating which components of terms are restricted to be monotonically non-decreasing.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

References

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* 11(1): 91–108.

See Also

addreg.smooth

---

addreg.smooth.design *Construct Design Matrix for addreg.smooth Model*

Description

addreg.smooth.design constructs the design matrix and associated formula for an addreg.smooth model, given a particular parameterisation, to be passed into addreg for fitting.

This is a workhorse function — it would not normally be called directly.

Usage

```
addreg.smooth.design(interpret, allref, design.knots, design.param)
```

Arguments

interpret	the object returned by running interpret.addreg.smooth for the desired model, containing details of the smooth components.
allref	the object returned by running addreg.smooth.allref for the desired model.
design.knots	a vector containing the number of internal knots for each smooth term (NA for Iso terms).
design.param	a vector of indices for a particular parameterisation in allref$ allref. Each element corresponds to a smooth term in the model, and the value indicates which item in the associated list is the reference level for this parameterisation.

Details

For a particular combination of reference levels, `addreg.smooth.design` constructs the associated design matrix and formula. Specifically, for `Iso` smooth components, it creates the matrix of indicator covariates for increments between levels. For `B` smooth components, it creates the basis functions using `splineDesign` and then either removing the column associated with the reference level, or transforming them into monotonic B-spline bases (see Donoghoe and Marschner, 2015).

The `formula` component is altered to include the terms in the design matrix, and `allref$monotonic` is altered such that all of the smooth coefficients are restricted to be non-negative, as required.

Value

A list with components:

`formula`	an updated version of `interpret$full.formula`, with smooth terms removed and replaced by the names of their associated basis components.
`data`	an updated version of `interpret$data`, with columns for the basis functions of the smooth terms added.
`monotonic`	an updated version of `allref$monotonic`, such that the coefficients associated with the smooth terms for this parameterisation are constrained to be non-negative.
`knots`	a list, with one component for each smooth term, containing the knot vector for that term (`NA` for `Iso` terms).

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

References

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* 11(1): 91–108.

See Also

`addreg.smooth`, `interpret.addreg.smooth`, `addreg.smooth.allref`, `addreg.design`

`anova.addreg` *Analysis of Deviance for addreg Fits*

Description

Compute an analysis of deviance table for more than one GLM fitted using `addreg`.

Usage

```
## S3 method for class 'addreg'
anova(object, ..., test = NULL)
```

Arguments

`object, ...`	objects of class `"addreg"`, typically the result of a call to `addreg`, or a `list` of `objects` for the "`addreglist`" method.
`test`	a character string, (partially) matching one of `"Chisq"`, `"LRT"`, `"Rao"`, `"F"` or `"Cp"`. See `stat.anova`.

Details

Unlike `anova.glm`, specifying a single object is not allowed.

The table has a row for the residual degrees of freedom and deviance for each model. For all but the first model, the change in degrees of freedom and deviance is also given. (This only makes statistical sense if the models are nested.) It is conventional to list the models from smallest to largest, but this is up to the user.

Models where the MLE lies on the boundary of the parameter space will be automatically removed from the list (with a warning), because asymptotic results to not apply to such models.

The table will optionally contain test statistics (and p-values) comparing the reduction in deviance for the row to the residuals. Mallows' *Cp* statistic is the residual deviance plus twice the estimate of σ^2 times the residual degrees of freedom, which is closely related to AIC. You can also choose `"LRT"` and `"Rao"` for likelihood ratio tests and Rao's efficient score test. The former is synonymous with `"Chisq"` (although both have an asymptotic chi-square distribution).

Value

An object of class `"anova"` inheriting from class `"data.frame"`.

Warning

The comparison between two or more models will only be valid if they are fitted to the same dataset. This may be a problem if there are missing values and R's default of `na.action = na.omit` is used, and `anova` will detect this with an error.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

See Also

addreg, anova.glm, anova

Examples

```
## For an example, see example(addreg)
```

B.Iso *Defining Smooths in addreg.smooth Formulae*

Description

Function used in the definition of smooth terms within addreg.smooth model formulae. The function does not evaluate a smooth — it exists purely to help set up a model using smooths.

Usage

```
B(..., knots = NULL, knot.range = 0:5)

Iso(...)
```

Arguments

...	variable that this smooth is a function of. Note that unlike gam, smooths that are functions of more than one variable are not supported.
knots	*unique* positions of *interior* knots of a B-spline basis. Boundary knots are created automatically.
knot.range	if knots is not specified, a vector containing a series of non-negative integers denoting the number of *interior* knots for which the model will be fit. These are placed at evenly-spaced quantiles of the observed covariate values. At least one of knots or knot.range must be non-missing.

Details

The function does not evaluate the variable arguments; the output from this function is passed as part of the arguments to addreg.smooth.design, which constructs the actual basis functions.

B is used to specify an order-3 B-spline basis (which can be restricted to be monotonically non-decreasing via the mono argument in addreg.smooth). If length(knot.

`range) > 1`, models with each of the specified number of interior knots will be fit, and the model with the best (smallest) `aic.c` will be returned.

`Iso` is used to specify an isotonic basis, designed such that the resulting function has non-negative increments at each observed covariate value. When `Iso` is used, the resulting function will always be monotonically non-decreasing, regardless of the value of `mono`.

Value

An object of class `"B.smooth"` (for `B`) or `"Iso.smooth"` (for `Iso`), which is a list with the following elements:

`term`	name of the term provided in the `...` argument.
`termlabel`	label for the term in the model; e.g. for term `"x"` it will be `"B(x)"` or `"Iso(x)"`.
`knots`	vector of interior knots (if specified). `NA` for `Iso`.
`knot.range`	vector of number of interior knots. `NA` for `Iso`.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`addreg.smooth`, `addreg.smooth.design`

`s` performs a similar function in the `mgcv` package.

Examples

```
## See example(addreg.smooth) for an example of specifying smooths in
##  model formulae.
```

`confint.addreg` *Confidence Intervals for addreg Model Parameters*

Description

Computes confidence intervals for one or more parameters in a fitted `addreg` model.

Usage

```
## S3 method for class 'addreg'
confint(object, parm, level = 0.95, ...)
```

Arguments

`object`	a fitted model object, resulting from a call to `addreg`.
`parm`	a specification of which parameters are to be given confidence intervals, either a vector of numbers or a vector of names. If missing, all parameters are considered.
`level`	the confidence level required.
`...`	additional argument(s) passed to `confint.default`.

Details

Calculates confidence intervals for model parameters assuming asymptotic normality, using `vcov.addreg(object)`. As such, if the MLE is on the boundary of the parameter space, (i.e. `object$boundary == TRUE`) the normality assumption is invalid and `NA` is returned.

Value

A matrix (or vector) with columns giving lower and upper confidence limits for each parameter. These will be labelled as (1-level)/2 and 1-(1-level)/2 in % (by default 2.5% and 97.5%).

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`confint.default`, `vcov.addreg`

Examples

```
## For an example, see example(addreg)
```

`contr.isotonic` *Contrast Matrix for Isotonic Covariate*

Description

Return something similar to a contrast matrix for a categorical covariate that we wish to be monotonically non-decreasing in a specified order.

Usage

```
contr.isotonic(n, perm, contrasts = TRUE, sparse = FALSE)
```

Arguments

`n`	a vector of levels for a factor, or the number of levels.
`perm`	a permutation of the levels of `n` (or of the numbers `1:n`), which define the order in which the coefficients must be monotonically non-decreasing.
`contrasts`	a logical indicating whether constrasts should be computed.
`sparse`	included for compatibility reasons. Has no effect.

Details

This function is used within `addreg.design` for categorical covariates with a specified order under a particular parameterisation. For Poisson and negative binomial models, this occurs if a categorical covariate is defined as monotonic; for binomial models, each parameterisation defines a permutation of the levels that must be monotonically increasing.

In the order specified by `perm`, the coefficient associated with each level is the sum of increments between the preceding levels. That is, the first level is defined as 0, the second as $0 + d_2$, the third as $0 + d_2 + d_3$, and so on. In fitting the model, these increments are constrained to be non-negative.

Note that these are not 'contrasts' as defined in the theory for linear models; rather this is used to define the `contrasts` attribute of each variable so that `model.matrix` produces the desired design matrix.

Value

A matrix with `n` rows and `k` columns, with `k=n-1` if `contrasts` is `TRUE` and `k=n` if `contrasts` is `FALSE`.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`addreg.design`, which uses `contr.isotonic` to create the design matrix using `model.matrix`.

`contr.treatment`, `contrasts` for their usual use in regression models.

Examples

```
contr.isotonic(4,1:4)
contr.isotonic(4,c(1,3,2,4))

# Show how contr.isotonic applies within model.matrix
x <- factor(round(runif(20,0,2)))
mf <- model.frame(~x)
contrasts(x) <- contr.isotonic(levels(x), levels(x))
model.matrix(mf)
```

`conv.test` *Convergence Test Based on L2 Norm*

Description

Performs a test of convergence based on the L2 norm of the change in the parameter estimates.

Usage

```
conv.test(theta1, theta2, epsilon)
```

Arguments

`theta1`	vector of parameter estimates at previous step.
`theta2`	vector of parameter estimates at current step.
`epsilon`	positive convergence tolerance.

Details

This is used as the convergence test in the `addreg` fitting functions, because the EM algorithm may converge slowly such that the test based on the deviance used in `glm.fit` (see `glm.control`) may report convergence at a point away from the actual optimum.

Value

A logical; `TRUE` if `sqrt(sum((theta1-theta2)**2))/sqrt(sum(theta1**2)) < epsilon`, `FALSE` otherwise.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

Examples

```
theta.old <- c(4,5,6)
theta.new <- c(4.05,5,6)

conv.test(theta.old, theta.new, 0.01)
conv.test(theta.old, theta.new, 0.005)
```

`interpret.addreg.smooth` *Interpret an addreg.smooth Formula*

Description

This is an internal function of package `addreg`. It is a service routine for `addreg.smooth` which interprets the smooth parts of the model formula and returns modified formulas to be used in the fitting functions.

Not normally called directly.

Usage

```
interpret.addreg.smooth(formula)
```

Arguments

`formula`	A formula as supplied to `addreg.smooth`, which includes at least one `B` or `Iso` term.

Value

A list with components:

`full.formula`	a `formula` object which is the same as the `formula` supplied, but with additional arguments removed from the smooth terms. E.g. `B(x, knot.range = 0:2)` would appear as `B(x)` in this formula.
`fake.formula`	a `formula` object which is the same as the `formula` supplied, but with smooth terms replaced by their covariates alone. E.g. `B(x, knot.range = 0:2)` would appear as `x` in this formula. Used to construct the model matrix.
`smooth.spec`	a named list containing the results of `eval`uating the smooth terms. See `B` and `Iso` for details.
`smooth.ind`	a vector containing the indices of the smooth components in the `formula`.
`terms`	the result of running `terms.formula(formula, specials = c("B", "Iso"))`.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

See Also

`addreg.smooth`

Examples

```
# Specify a smooth model with knot.range
res <- interpret.addreg.smooth(y ~ B(x, knot.range = 0:2) + x2)
# The knot.range is removed from the full.formula...
print(res$full.formula)
# ...but is stored in the $smooth.spec component of the result:
print(res$smooth.spec$x$knot.range)
```

---

negbin1 *Family Functions for Negative Binomial 1 Models*

Description

Specifies the information required to fit a negative binomial 1 (NB1) model.

Usage

```
negbin1(link, phi = stop("'phi' must be given"))
```

Arguments

`link`	included for compatibility with `family`. For `addreg` models, this will always be `"identity"`.
`phi`	the value of the scale parameter of the NB1 distribution (see "Details"). This can be set to `NA` for initialisation, but during estimation the family should be updated with the current estimate, and must be strictly positive.

Details

The NB1 distribution can be parameterised in terms of a mean μ and scale parameter ϕ (the `phi` argument of this function), such that if $Y \sim NB1(\mu, \phi)$, then $E(Y) = \mu$ and $Var(Y) = (1 + \phi)\mu$.

These can be related to the `size` and `prob` arguments of the `NegBinomial` functions by `size` $= \mu/\phi$ and `prob` $= 1/(1 + \phi)$.

Value

An object of class `"family"`: see `family` for full details. Note that when the estimate of `phi` is updated in a model, this `family` object must be reloaded using the new estimate.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

nnnegbin	*ECME Algorithm for Additive Negative Binomial 1 Model*

Description

Finds the maximum likelihood estimate of an additive negative binomial (NB1) model using an ECME algorithm, where each of the mean coefficients is restricted to be non-negative.

Usage

```
nnnegbin(y, x, standard, offset, start, control = list())
```

Arguments

y	non-negative integer response vector.
x	non-negative covariate matrix.
standard	standardising vector, where each element is a positive constant that (multiplicatively) standardises the fitted value of the corresponding element of the response vector. The default is a vector of ones.
offset	non-negative additive offset vector. The default is a vector of zeros.
start	vector of starting values for the parameter estimates. The last element is the starting value of the `scale`, and must be `> 1`. The remaining elements are for the additive mean parameters, and must be greater than `control$bound.tol`.
control	an `addreg.control` object, which controls the fitting process.

Details

This is a workhorse function for `addreg`, and runs the ECME algorithm to find the constrained non-negative MLE associated with an additive NB1 model.

Value

A list containing the following components

coefficients	the constrained non-negative maximum likelihood estimate of the mean parameters.
scale	the maximum likelihood estimate of the scale parameter.
residuals	the residuals at the MLE, that is `y - fitted.values`
fitted.values	the fitted mean values.

`rank`	the number of parameters in the model (named "`rank`" for compatibility — we assume that models have full rank)
`family`	included for compatibility — will always be `negbin1(identity)`.
`linear.predictors`	included for compatibility — same as `fitted.values` (as this is an identity-link model).
`deviance`	up to a constant, minus twice the maximised log-likelihood (with respect to a saturated NB1 model with the same `scale`).
`aic`	a version of Akaike's *An Information Criterion*, minus twice the maximised log-likelihood plus twice the number of parameters.
`aic.c`	a small-sample corrected version of Akaike's *An Information Criterion* (Hurvich, Simonoff and Tsai, 1998).
`null.deviance`	the deviance for the null model, comparable with `deviance`. The null model will include the offset and an intercept.
`iter`	the number of iterations of the EM algorithm used.
`weights`	included for compatibility — a vector of ones.
`prior.weights`	included for compatibility — a vector of ones.
`standard`	the `standard` vector passed to this function.
`df.residual`	the residual degrees of freedom.
`df.null`	the residual degrees of freedom for the null model.
`y`	the `y` vector used.
`converged`	logical. Did the ECME algorithm converge (according to `conv.test`)?
`boundary`	logical. Is the MLE on the boundary of the parameter space — i.e. are any of the `coefficients < control$bound.tol`?
`loglik`	the maximised log-likelihood.
`nn.design`	the non-negative `x` matrix used.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>.

References

Hurvich, C. M., J. S. Simonoff and C.-L. Tsai (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(2): 271–293.

`nnpois` *EM Algorithm for Identity-link Poisson GLM*

Description

Finds the maximum likelihood estimate of an identity-link Poisson GLM using an EM algorithm, where each of the coefficients is restricted to be non-negative.

Usage

```
nnpois(y, x, standard, offset, start, control = list())
```

Arguments

`y`	non-negative integer response vector.
`x`	non-negative covariate matrix.
`standard`	standardising vector, where each element is a positive constant that (multiplicatively) standardises the fitted value of the corresponding element of the response vector. The default is a vector of ones.
`offset`	non-negative additive offset vector. The default is a vector of zeros.
`start`	starting values for the parameter estimates. Each element must be greater than `control$bound.tol`.
`control`	an `addreg.control` object, which controls the fitting process.

Details

This is a workhorse function for `addreg`, and runs the EM algorithm to find the constrained non-negative MLE associated with an identity-link Poisson GLM. See Marschner (2010) for full details.

Value

A list containing the following components

`coefficients`	the constrained non-negative maximum likelihood estimate of the parameters.
`residuals`	the residuals at the MLE, that is `y - fitted.values`
`fitted.values`	the fitted mean values.
`rank`	the number of parameters in the model (named "`rank`" for compatibility — we assume that models have full rank)
`family`	included for compatibility — will always be `poisson(identity)`.

`linear.predictors`
: included for compatibility — same as `fitted.values` (as this is an identity-link model).

`deviance`
: up to a constant, minus twice the maximised log-likelihood.

`aic`
: a version of Akaike's *An Information Criterion*, minus twice the maximised log-likelihood plus twice the number of parameters.

`aic.c`
: a small-sample corrected version of Akaike's *An Information Criterion* (Hurvich, Simonoff and Tsai, 1998).

`null.deviance`
: the deviance for the null model, comparable with `deviance`. The null model will include the offset and an intercept.

`iter`
: the number of iterations of the EM algorithm used.

`weights`
: included for compatibility — a vector of ones.

`prior.weights`
: included for compatibility — a vector of ones.

`standard`
: the `standard` vector passed to this function.

`df.residual`
: the residual degrees of freedom.

`df.null`
: the residual degrees of freedom for the null model.

`y`
: the `y` vector used.

`converged`
: logical. Did the EM algorithm converge (according to `conv.test`)?

`boundary`
: logical. Is the MLE on the boundary of the parameter space — i.e. are any of the `coefficients < control$bound.tol`?

`loglik`
: the maximised log-likelihood.

`nn.design`
: the non-negative `x` matrix used.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>.

This function is based on code from Marschner, Gillett and O'Connell (2012) written by Alexandra Gillett.

References

Hurvich, C. M., J. S. Simonoff and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(2): 271–293.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* 19(3): 666–683.

Marschner, I. C., A. C. Gillett and R. L. O'Connell (2012). Stratified additive Poisson models: Computational methods and applications in clinical epidemiology. *Computational Statistics and Data Analysis* 56(5): 1115–1130.

`plot.addreg.smooth` *Default addreg.smooth Plotting*

Description

Takes a fitted `addreg.smooth` object produced by `addreg.smooth` and plots the component smooth functions that make it up, on the scale of the linear predictor, for specified values of the other covariates.

Usage

```
## S3 method for class 'addreg.smooth'
plot(x, type = c("response", "link"), at = data.frame(),
     knotlines = TRUE, nobs = 1000, ...)
```

Arguments

`x`	a fitted `addreg.smooth` object as produced by `addreg.smooth`.
`type`	the type of prediction required. Note that, unlike `predict.addreg.smooth`, `"terms"` is not a valid option. Also, because `addreg.smooth` only applies identity-link models, `"response"` and `"link"` will have the same results — they are included for consistency.
`at`	a data frame containing the values at which the prediction should be evaluated. The columns must contain the covariates in the model, and several rows may be provided (in which case, multiple lines are drawn on the same plot). Cannot be missing or `NULL`.
`knotlines`	logical; if vertical lines should be drawn on the plot to indicate the locations of the knots for B-spline terms.
`nobs`	the number of points which should be used to create the curve. These are placed evenly along the range of the observed covariate values from the original model.
`...`	other graphics parameters to pass on to plotting commands (note: some will not work).

Details

For each smooth covariate in the model of `x`, `predict.addreg.smooth` is used to obtain predicted values for the range of that covariate, with the other covariates remaining fixed at their values given in `at`. Several rows may be provided in `at`, in which case, one curve is drawn for each, and they are coloured using `rainbow(nrow(at))`. If the model contains a single smooth covariate and no other covariates, `at` may be provided as an empty data frame, `data.frame()`.

Value

The function simply generates plots.

Note

If this function is too restrictive, it may be easier to use `predict.addreg.smooth` to get predictions for the dataset of your choice, and do the plotting manually.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`addreg.smooth`, `predict.addreg.smooth`

Examples

```
## For an example, see example(addreg.smooth)
```

`predict.addreg` *Predict Method for addreg Fits*

Description

Obtains predictions from a fitted `addreg` object.

Usage

```
## S3 method for class 'addreg'
predict(object, newdata = NULL, type = c("link", "response",
        "terms"), terms = NULL, na.action = na.pass,
        checkminmax = TRUE, ...)
```

Arguments

`object`	a fitted object of class inheriting from `"addreg"`.
`newdata`	optionally, a data frame in which to look for variables with which to predict. If omitted, the fitted linear predictors are used.
`type`	the type of prediction required. The default is on the scale of the linear predictors; the alternative `"response"` is on the scale of the response variable. The `"terms"` option returns a matrix giving the fitted values of each term in the model formula on the linear predictor scale. The value of this argument can be abbreviated.
`terms`	with `type = "terms"` by default all terms are returned. A character vector specifies which terms are to be returned.

na.action	function determining what should be done with missing values in newdata. The default is to predict NA.
checkminmax	logical indicating whether or not values of continuous covariates in newdata should be checked to ensure they lie within the covariate space associated with the fitted model. Otherwise predicted values could lie outside the parameter space.
...	further arguments passed to or from other methods.

Details

If newdata is omitted the predictions are based on the data used for the fit. In that case how cases with missing values in the original fit are treated is determined by the na.action argument of that fit. If na.action = na.omit, omitted cases will not appear in the residuals; if na.action = na.exclude they will appear, with residual value NA. See also napredict.

Value

A vector or matrix of predictions. For type = "terms", this is a matrix with a column per term, and may have an attribute "constant".

Note

Variables are first looked for in newdata and then searched for in the usual way (which will include the environment of the formula used in the fit). A warning will be given if the variables found are not of the same length as those in newdata if it was supplied.

Author(s)

Mark W. Donoghoe <Mark.Donoghoe@mq.edu.au>

See Also

addreg

predict.glm for the equivalent method for models fit using glm.

Examples

```
## For an example, see example(addreg)
```

`predict.addreg.smooth` *Predict Method for addreg.smooth Fits*

Description

Obtains predictions from a fitted `addreg.smooth` object.

Usage

```
## S3 method for class 'addreg.smooth'
predict(object, newdata = NULL, type = c("link", "response",
        "terms"), terms = NULL, na.action = na.pass, ...)
```

Arguments

`object`	a fitted object of class inheriting from `"addreg.smooth"`.
`newdata`	optionally, a data frame in which to look for variables with which to predict. If omitted, the fitted linear predictors are used.
`type`	the type of prediction required. The default is on the scale of the linear predictors; the alternative `"response"` is on the scale of the response variable. The `"terms"` option returns a matrix giving the fitted values of each term in the model formula on the linear predictor scale. The value of this argument can be abbreviated.
`terms`	with `type = "terms"` by default all terms are returned. A character vector specifies which terms are to be returned.
`na.action`	function determining what should be done with missing values in `newdata`. The default is to predict `NA`.
`...`	further arguments passed to or from other methods.

Details

`predict.addreg.smooth` constructs the underlying basis functions for smooth variables in `newdata` and runs `predict.addreg` to obtain predictions. Note that if values of smooth covariates in `newdata` are outside the covariate space of `object`, an error will be returned.

If `newdata` is omitted, the predictions are based on the data used for the fit. In that case how cases with missing values in the original fit are treated is determined by the `na.action` argument of that fit. If `na.action = na.omit`, omitted cases will not appear in the residuals; if `na.action = na.exclude` they will appear, with residual value `NA`. See also `napredict`.

Value

A vector or matrix of predictions. For `type = "terms"`, this is a matrix with a column per term, and may have an attribute `"constant"`.

Note

Variables are first looked for in `newdata` and then searched for in the usual way (which will include the environment of the formula used in the fit). A warning will be given if the variables found are not of the same length as those in `newdata` if it was supplied.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`addreg.smooth`, `predict.addreg`

`predict.glm` for the equivalent method for models fit using `glm`.

Examples

```
## For an example, see example(addreg.smooth)
```

`summary.addreg` *Summarizing addreg Model Fits*

Description

These functions are all `methods` for class `addreg` or `summary.addreg` objects.

Usage

```
## S3 method for class 'addreg'
summary(object, correlation = FALSE, ...)

## S3 method for class 'summary.addreg'
print(x, digits = max(3L, getOption("digits") - 3L),
      signif.stars = getOption("show.signif.stars"), ...)
```

Arguments

`object`	an object of class `"addreg"`, usually from a call to `addreg` or `addreg.smooth`.
`x`	an object of class `"summary.addreg"`, usually from a call to `summary.addreg`.
`correlation`	logical; if `TRUE`, the correlation matrix of the estimated parameters is returned and printed.
`digits`	the number of significant digits to use when printing.
`signif.stars`	logical; if `TRUE`, 'significance stars' are printed for each coefficient.
`...`	further arguments passed to or from other methods.

Details

These perform the same function as `summary.glm` and `print.summary.glm`, producing similar results for `addreg` models. `print.summary.addreg` additionally prints the small-sample corrected AIC (`aic.c`), the number of EM iterations for the parameterisation corresponding to the MLE, and for negative binomial models, the estimate of ϕ (`scale`-1) and its standard error.

The dispersion used in calculating standard errors is fixed as 1 for binomial and Poisson models, and is estimated via maximum likelihood for negative binomial models.

Value

`summary.addreg` returns an object of class `"summary.addreg"`, a list with components

`call`	the component from `object`.
`family`	the component from `object`.
`deviance`	the component from `object`.
`aic`	the component from `object`.
`aic.c`	the component from `object`.
`df.residual`	the component from `object`.
`null.deviance`	the component from `object`.
`df.null`	the component from `object`.
`iter`	the component from `object`.
`deviance.resid`	the deviance residuals: see `residuals.glm`.
`coefficients`	the matrix of coefficients, standard errors, z-values and p-values.
`aliased`	included for compatibility — always `FALSE`.
`dispersion`	the inferred/estimated dispersion.
`df`	included for compatibility — a 3-vector of the number of coefficients, the number of residual degrees of freedom, and the number of coefficients (again).
`cov.unscaled`	the unscaled (`dispersion = 1`) estimated covariance matrix of the estimated coefficients. `NaN` if `object$boundary == TRUE`.
`cov.scaled`	ditto, scaled by `dispersion`.
`correlation`	if `correlation` is `TRUE`, the estimated correlations of the estimated coefficients. `NaN` if `object$boundary == TRUE`.

For negative binomial models, the object also contains

`phi`	the estimate of ϕ (`scale`-1).
`var.phi`	the estimated variance of `phi`.

Note

If `object$boundary == TRUE`, the standard errors of the coefficients are not valid, and a matrix of `NaN`s is returned by `vcov.addreg`.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`addreg`, `summary.glm`

Examples

```
## For an example, see example(addreg)
```

`vcov.addreg` *Calculate Variance-Covariance Matrix for a Fitted addreg Model Object*

Description

Returns the variance-covariance matrix of the main parameters of a fitted `addreg` model object.

Usage

```
## S3 method for class 'addreg'
vcov(object, ...)
```

Arguments

`object`	an object of class `"addreg"`, usually from a call to `addreg` or `addreg.smooth`.
`...`	additional arguments for method functions.

Details

An equivalent method to `vcov`, to use with `addreg` models.

Value

A matrix of the estimated covariances between the parameter estimates in the linear or non-linear predictor of the model. This should have row and column names corresponding to the parameter names given by the `coef` method.

Note

If `object$boundary == TRUE`, the standard errors of the coefficients are not valid, and a matrix of `NaN`s is returned.

Author(s)

Mark W. Donoghoe <`Mark.Donoghoe@mq.edu.au`>

See Also

`summary.addreg`, `vcov.glm`

Examples

```
## For an example, see example(addreg)
```

C

Conference proceedings paper

This appendix contains a conference proceedings paper that was presented as an oral presentation at the 29th International Workshop on Statistical Modelling, held 14–18 July 2014 at Georg-August-Universität Göttingen, Germany. The theoretical content is an earlier, condensed version of the method for semi-parametric regression described in detail in Chapter 4 of this thesis, but the paper demonstrates an application of the method to a different dataset, from the BOOST-NZ study (see Section 1.4.3).

The citation for the proceedings paper is:

Donoghoe, M. W. and I. C. Marschner. Smooth semi-parametric adjustment of rate differences, risk differences and relative risks. *Proceedings of the 29th International Workshop on Statistical Modelling*. Ed. by T. Kneib, F. Sobotka, J. Fahrenholz, and H. Irmer, 2014. **1**: 105–110.

Smooth semi-parametric adjustment of rate differences, risk differences and relative risks

Mark W. Donoghoe[1,2], Ian C. Marschner[1,2]

[1] Department of Statistics, Macquarie University, NSW 2109, Australia

[2] NHMRC Clinical Trials Centre, University of Sydney, NSW 2006, Australia

Abstract

New computational methods have recently been developed that allow stable fitting of constrained GLMs with bounded non-canonical link functions, such as the log-link binomial model. By employing B-splines, we can extend these approaches to allow for semi-parametric adjustment of rate differences, risk differences and relative risks. These methods provide alternatives to standard fitting methods, resulting in greater stability for accommodating the required parameter bounds. They also provide a straightforward way to accommodate additional restrictions such as monotonic regression functions. We demonstrate an application to data from a clinical trial of oxygen supplementation in premature infants.

Keywords: Generalised additive model · Semi-parametric model · Rate difference · Risk difference · Relative risk

C.1 Introduction

Rate differences, risk differences and relative risks are often useful effect measures in biostatistical settings, and their analogues also have broad applicability in other areas of statistics. However, in order to adjust for covariates we must use a constrained generalised linear model (GLM) with a non-canonical link where the fitted means are restricted to a bounded interval. These GLMs include the log-link binomial, and identity-link Poisson and binomial models. Common fitting methods based on Fisher scoring and other Newton-type algorithms can fail to converge to the maximum likelihood estimate (MLE) in this situation.

It is therefore useful to have more stable methods for fitting these models. Combinatorial EM (CEM) algorithms have recently been developed for these GLMs, allowing stable computation of the MLE. Using B-splines, we extend these methods to generalised additive models (GAMs), where continuous covariates can have a semi-parametric relationship with the outcome. This approach leads to greater stability for accommodating the required parameter bounds, and allows additional model constraints such as monotonic regression functions.

C.2 Method

The GLM with link function g is extended to a GAM by the introduction of C continuous covariates that affect $g(\mu)$ through the unspecified functions $f_1, \ldots, f_C$. We restrict our estimate of each f_c to the space defined by a chosen set of basis functions, such that

$$f_c(w) = \sum_{d=1}^{D_c} \gamma_{cd} B_{cd}(w).$$

The basis functions we use here are the B-splines of order 3, which are strictly non-negative. Thus if all of the coefficients are non-negative, $f_c(w)$ will be non-negative for all w; and likewise if the coefficients are non-positive, the curve will always be non-positive. The B-splines are normalised such that $\sum_d B_{cd}(w) = 1$, which means that we must apply an identifiability constraint $\gamma_{ct_c} = 0$ for some t_c.

When $C = 0$, methods have been developed for estimating the MLE for identity-link Poisson (Marschner, 2010), log-link binomial (Marschner and Gillett, 2012) and identity-link binomial GLMs. The methods are all CEM algorithms (Marschner, 2014), which will always converge to the MLE. With these methods, we are also able to restrict certain coefficients to be non-negative or non-positive.

CEM algorithms require that the parameter space is partitioned into distinct subspaces, and use an EM algorithm to find the constrained MLE within each. One of these constrained MLEs will be the overall MLE. For these GAMs, we partition the parameter space based on the index of the smallest or largest B-spline coefficient, which can be achieved by setting a particular $\gamma_{ct_c} = 0$ and restricting the remaining coefficients to be non-negative or non-positive. We repeat this process for all possible choices of t_c and find the constrained MLE for each, one of which will coincide with the overall MLE.

A sufficient condition for f_c to be monotonically non-decreasing is that the sequence of B-spline coefficients is non-decreasing. To apply a monotonicity constraint to any of these models, we can reparameterise the smooth curve such that we are estimating the increments between successive coefficients, and can constrain these to be non-negative or non-positive, as required.

C.3 Application

The BOOST-NZ study (Darlow et al., 2014) was a randomised trial in premature infants, comparing the effects of different target ranges for oxygen saturation (SpO_2). Both high and low levels of oxygen are associated with mortality and other complications, so the primary outcome of the study was death or major disability at two years of age. Unadjusted analysis of the primary outcome showed a relative risk of 1.16 (95% CI 0.90–1.50) and a risk difference of 0.06 (95% CI −0.04–0.17), with lower risk in the low-target group.

We use the methods outlined in Section C.2 to adjust these effect measures for the actual level of oxygen that the infant received. Each infant's median SpO_2 level while receiving supplementary oxygen was entered as the semi-parametric covariate into each model, and the results are shown in Figure C.1.

The adjusted analyses show that the minimum risk is associated with an SpO_2 close to 94%. The adjusted effect of randomised treatment is a relative risk of 1.46 (95% CI 1.04–2.07) and a risk difference of 0.19 (95% CI 0.06–0.33). The confidence intervals for these parameters were estimated using a normal approximation.

For the outcome of mortality, the risk of death decreases as the SpO_2 level increases, and so we can restrict the semi-parametric curve to be monotonically non-increasing. The unadjusted effect of treatment is a relative risk of 1.08 (95% CI 0.65–1.78) or a risk difference of 0.01 (95% CI −0.06–0.09) in favour of the low-target group.

The results of the adjusted analyses are shown in Figure C.2. The adjusted relative risk is estimated to be 1.89, and the adjusted risk difference is 0.04.

The estimates from these models are on the boundary of their respective parameter

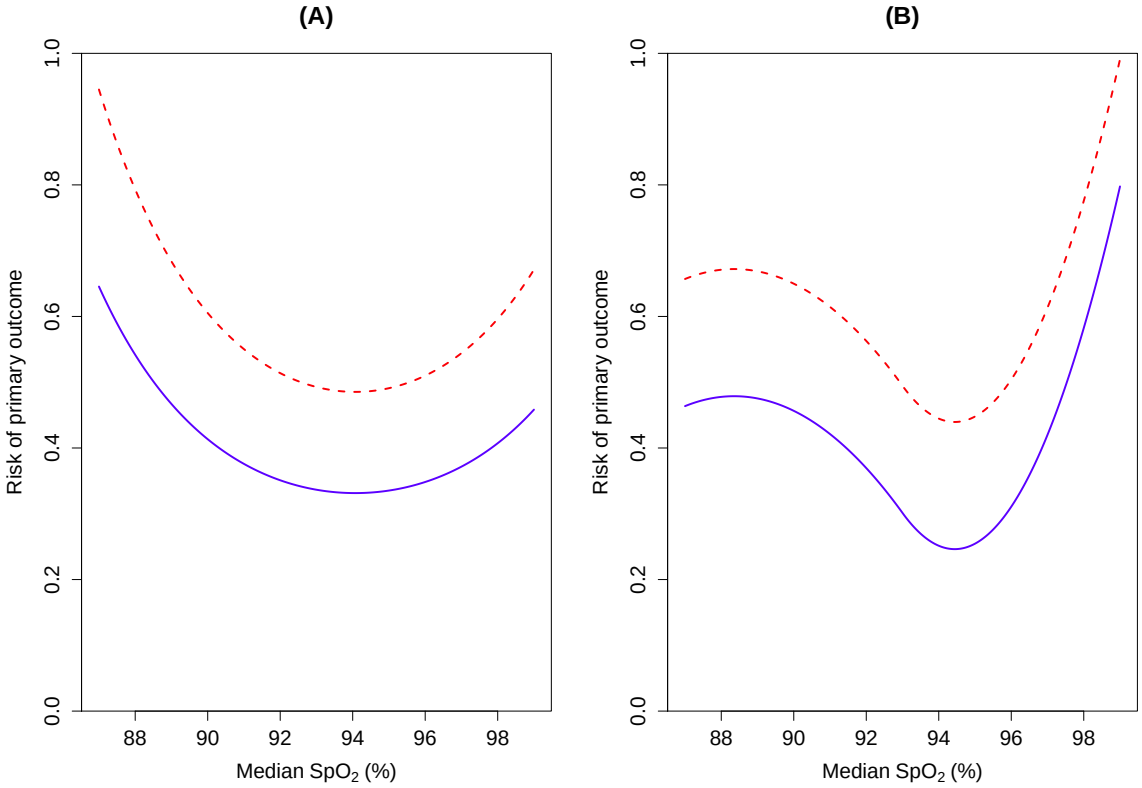


Figure C.1: Risk of primary outcome by median SpO_2 and randomised treatment (blue solid = low target, red dashed = high target) in BOOST-NZ, under (A) log-link and (B) identity-link binomial models.

spaces, so we must estimate confidence intervals using bootstrap resampling. Importantly, the algorithm will converge to the MLE in every bootstrap sample, eliminating bias due to non-convergence. From 1000 bootstrap samples, we estimate the 95% confidence intervals to be 1.10–2.86 for the relative risk, and −0.03–0.10 for the risk difference.

C.4 Other methods

We compared our approach with other methods for fitting GAMs that have been implemented in R, and were able to show that our method has advantages over existing methods in some contexts.

The most notable existing methods are implemented in the `gam` (Hastie, 2013), `mgcv` (Wood, 2011) and `gamlss` (Rigby and Stasinopoulos, 2005) packages. The fitting procedures underlying these approaches each employ a Newton-type algorithm, which

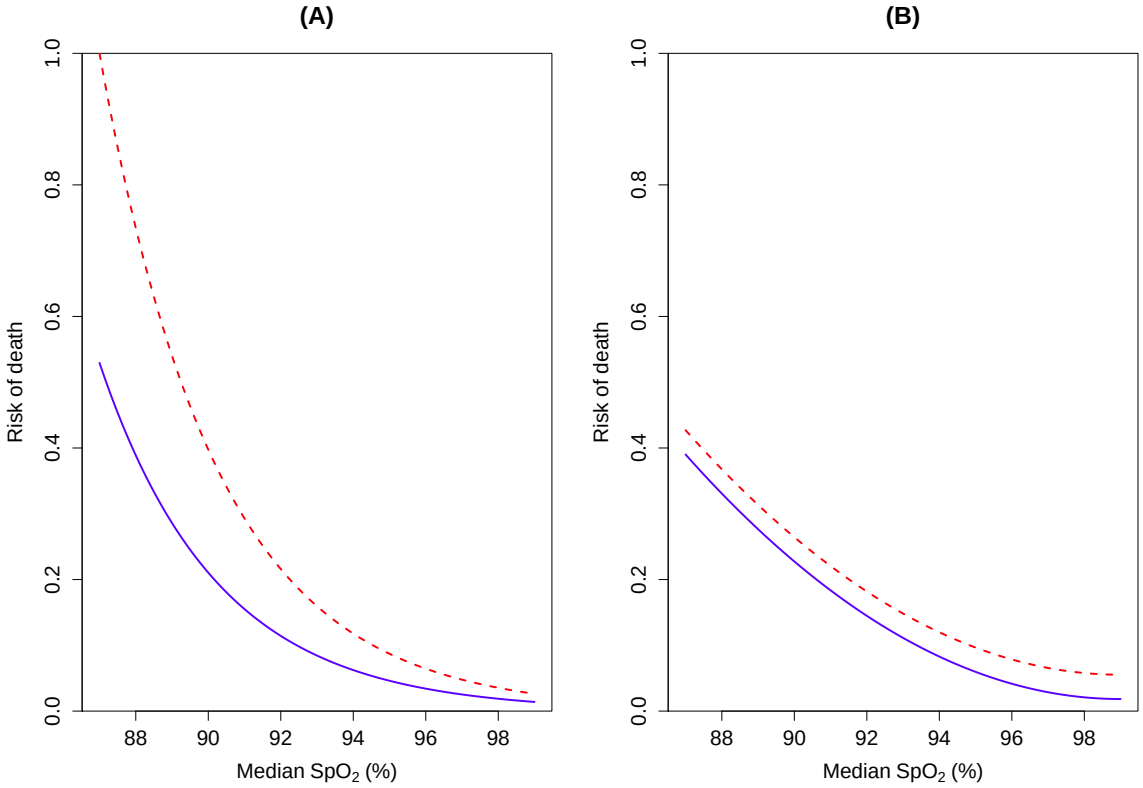


FIGURE C.2: Risk of death by median SpO_2 and randomised treatment (blue solid = low target, red dashed = high target) in BOOST-NZ, under (A) log-link and (B) identity-link binomial models.

is not guaranteed to converge to the MLE unless step-size optimisation is performed.

In fact, of these packages, only `mgcv` incorporates automatic step-halving if the potential update of the parameter estimates moves outside the parameter space. This method reported convergence in all 1000 bootstrap samples for the analysis in Figure C.1, for both the log and identity links. However, in some cases `mgcv` converged to sub-optimal parameter estimates, particularly when the MLE was on the boundary of the parameter space. Furthermore, `mgcv` is unable to accommodate the monotonicity constraint for the analysis depicted in Figure C.2.

The `gamlss` package allows the user to specify the step size for updating the parameter estimates and also offers the option to use step-halving if the deviance increases at a particular iteration. However, the method terminates with an error if the parameter estimates move outside the parameter space, making it inappropriate for automated model-fitting such as bootstrapping. It failed to converge in 52 of the 1000 bootstrap samples using the log link, and did not converge in any of the samples when we used

the identity link.

The `gam` package does not include either step-halving or any check for the validity of the parameter estimates. As such, it may fail to converge or converge to a solution outside the parameter space, which occurred in 844 and 963 of the 1000 bootstrap samples, using the log and identity links respectively.

A difference with these methods is that they maximise a penalised likelihood, allowing greater flexibility in the number and positioning of the knots while discouraging large fluctuations in the resulting smooth curve. Penalised likelihood could be incorporated into our methods by a similar approach to that used by Marschner and Gillett (2012, Supplementary materials), but this would add substantially to the computational load.

Aside from its stability, another benefit of our approach is that it is straightforward to impose monotonicity constraints on selected smooth curves. If it is appropriate to assume monotonicity, this can reduce the spurious fluctuations in the estimated curve, and possibly increase the efficiency of the parameter estimates in the model.

The `GMBBoost` (Leitenstorfer and Tutz, 2007) and `GMonBoost` (Tutz and Leitenstorfer, 2007) functions employ the technique of likelihood boosting to apply a monotonicity constraint to smooth functions in maximising a penalised log-likelihood. The current implementation of both, however, only allows canonical link functions, and therefore cannot be used to fit the models considered in this paper.

Acknowledgements

Special thanks to Prof. Brian Darlow, University of Otago, New Zealand for permission to present the BOOST-NZ data. This research was supported by the Australian Research Council (DP110101254).

References

Darlow, B. A., S. L. Marschner, M. Donoghoe, M. R. Battin, R. S. Broadbent, M. J. Elder, M. P. Hewson, M. P. Meyer, A. Ghadge, P. Graham, N. J. McNeill, C. A. Kuschel, and W. O. Tarnow-Mordi (2014). Randomized controlled trial of oxygen saturation targets in very preterm infants: Two year outcomes. *The Journal of*

Pediatrics **165**(1): 30–35.e2. DOI: 10.1016/j.jpeds.2014.01.017.

Hastie, T. J. (2013). *gam: Generalized Additive Models*. R package version 1.09. URL: http://CRAN.R-project.org/package=gam.

Leitenstorfer, F. and G. Tutz (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8**(3): 654–673. DOI: 10.1093/biostatistics/kxl036.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**(3): 666–683. DOI: 10.1198/jcgs.2010.09127.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* **24**(6): 921–940. DOI: 10.1007/s11222-013-9411-7.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics* **13**(1): 179–192. DOI: 10.1093/biostatistics/kxr030.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**: 507–554. DOI: 10.1111/j.1467-9876.2005.00510.x.

Tutz, G. and F. Leitenstorfer (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics* **16**(1): 165–188. DOI: 10.1198/106186007X180949.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(1): 3–36. DOI: 10.1111/j.1467-9868.2010.00749.x.

Bibliography

Adelstein, B.-A., T. A. Dobbins, C. A. Harris, I. C. Marschner, and R. L. Ward (2011). A systematic review and meta-analysis of KRAS status as the determinant of response to anti-EGFR antibodies and the impact of partner chemotherapy in metastatic colorectal cancer. *European Journal of Cancer* **47**(9): 1343–1354. DOI: 10.1016/j.ejca.2011.03.031.

Aiello, L. M. (2003). Perspectives on diabetic retinopathy. *American Journal of Ophthalmology* **136**(1): 122–135. DOI: 10.1016/S0002-9394(03)00219-8.

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**(3): 251–262. DOI: 10.1007/bf00140869.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716–723. DOI: 10.1109/tac.1974.1100705.

Armitage, P. (1957). Studies in the variability of pock counts. *The Journal of Hygiene* **55**(4): 564–581. DOI: 10.1017/S0022172400037451.

ASSENT-2 Investigators (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *The Lancet* **354**(9180): 716–722. DOI: 10.1016/S0140-6736(99)07403-6.

ASSENT-3 Investigators (2001). Efficacy and safety of tenecteplase in combination with enoxaparin, abciximab, or unfractionated heparin: The ASSENT-3 randomised trial in acute myocardial infarction. *The Lancet* **358**(9282): 605–613. DOI: 10.1016/S0140-6736(01)05775-0.

Baker, S. G. (1994). The multinomial–Poisson transformation. *The Statistician* **43**(4): 495–504. DOI: 10.2307/2348134.

Bates, G. E. and J. Neyman (1952). Contributions to the theory of accident proneness. II: True or false contagion. *University of California Publications in Statistics* **1**(10): 255–276.

Behr, A. and F. Heid (2011). The success of bank mergers revisited. An assessment based on a matching strategy. *Journal of Empirical Finance* **18**(1): 117–135. DOI: 10.1016/j.jempfin.2010.08.006.

Berger, D., A. Corvalan, W. Easterly, and S. Satyanath (2013). Do superpower interventions have short and long term consequences for democracy? *Journal of Comparative Economics* **41**(1): 22–34. DOI: 10.1016/j.jce.2013.01.004.

Berry, D. C., P. Knapp, and T. Raynor (2006). Expressing medicine side effects: Assessing the effectiveness of absolute risk, relative risk, and number needed to harm, and the provision of baseline risk information. *Patient Education and Counseling* **63**(1-2): 89–96. DOI: 10.1016/j.pec.2005.09.003.

Blizzard, L. and D. W. Hosmer (2006). Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal* **48**(1): 5–22. DOI: 10.1002/bimj.200410165.

Bobbio, M., B. Demichelis, and G. Giustetto (1994). Completeness of reporting trial results: Effect on physicians' willingness to prescribe. *The Lancet* **343**(8907): 1209–1211. DOI: 10.1016/S0140-6736(94)92407-4.

Böhning, D. and J. Sarol (2000). Estimating risk difference in multicenter studies under baseline-risk heterogeneity. *Biometrics* **56**(1): 304–308. DOI: 10.1111/j.0006-341X.2000.00304.x.

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer-Verlag.

Cameron, A. C. and P. Johansson (1997). Count data regression using series expansions: With applications. *Journal of Applied Econometrics* **12**(3): 203–223. DOI: 10.1002/(SICI)1099-1255(199705)12:3<203::AID-JAE446>3.0.CO;2-2.

Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* **1**(1): 29–53. DOI: 10.1002/jae.3950010104.

Cameron, A. C. and P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

Cheung, Y. B. (2007). A modified least-squares regression approach to the estimation of risk difference. *American Journal of Epidemiology* **166**(11): 1337–1344. DOI: 10.1093/aje/kwm223.

Cochran, W. G. (1954). Some methods for strengthening the common Chi-squared tests. *Biometrics* **10**(4): 417–451. DOI: 10.2307/3001616.

Consul, P. C. (1989). *Generalized Poisson Distributions: Properties and Applications*. New York: Marcel Dekker Inc.

Cook, R. J. and D. L. Sackett (1995). The number needed to treat: A clinically useful measure of treatment effect. *BMJ* **310**(6977): 452–454. DOI: 10.1136/bmj.310.6977.452.

Cook, T. D. (2002). Advanced Statistics: Up with odds ratios! A case for odds ratios when outcomes are common. *Academic Emergency Medicine* **9**(12): 1430–1434. DOI: 10.1197/aemj.9.12.1430.

Cummings, P. (2009). The relative merits of risk ratios and odds ratios. *Archives of Pediatrics and Adolescent Medicine* **163**(5): 438–445. DOI: 10.1001/archpediatrics.2009.31.

Darlow, B. A., S. L. Marschner, M. Donoghoe, M. R. Battin, R. S. Broadbent, M. J. Elder, M. P. Hewson, M. P. Meyer, A. Ghadge, P. Graham, N. J. McNeill, C. A. Kuschel, and W. O. Tarnow-Mordi (2014). Randomized controlled trial of oxygen saturation targets in very preterm infants: Two year outcomes. *The Journal of Pediatrics* **165**(1): 30–35.e2. DOI: 10.1016/j.jpeds.2014.01.017.

Davies, H. T. O., I. K. Crombie, and M. Tavakoli (1998). When can odds ratios mislead? *BMJ* **316**(7136): 989–991. DOI: 10.1136/bmj.316.7136.989.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Dean, C., J. F. Lawless, and G. E. Willmot (1989). A mixed Poisson–inverse-Gaussian regression model. *The Canadian Journal of Statistics* **17**(2): 171–181. DOI: 10.2307/3314846.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1): 1–38.

Donoghoe, M. W. (2015a). *addreg: Additive Regression for Discrete Data*. R package version 2.0. URL: http://CRAN.R-project.org/package=addreg.

Donoghoe, M. W. (2015b). *logbin: Relative Risk Regression Using the Log-Binomial Model*. R package version 1.2. URL: http://CRAN.R-project.org/package=logbin.

Donoghoe, M. W. and I. C. Marschner (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis* **80**: 184–196. DOI: 10.1016/j.csda.2014.06.019.

Donoghoe, M. W. and I. C. Marschner (2015). Flexible regression models for rate differences, risk differences and relative risks. *International Journal of Biostatistics* **11**(1): 91–108. DOI: 10.1515/ijb-2014-0044.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* **81**(395): 709–721. DOI: 10.2307/2289002.

Egger, M., G. D. Smith, and A. N. Phillips (1997). Meta-analysis: Principles and procedures. *BMJ: British Medical Journal* **315**(7121): 1533–1537. DOI: 10.1136/bmj.315.7121.1533.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**(2): 89–121. DOI: 10.1214/ss/1038425655.

Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statistica Neerlandica* **38**(3): 159–167. DOI: 10.1111/j.1467-9574.1984.tb01107.x.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13**(1): 342–368. DOI: 10.1214/aos/1176346597.

FIELD Study Investigators (2005). Effects of long-term fenofibrate therapy on cardiovascular events in 9795 people with type 2 diabetes mellitus (the FIELD study): Randomised controlled trial. *The Lancet* **366**(9500): 1849–1861. DOI: 10.1016/S0140-6736(05)67667-2.

Forrow, L., W. C. Taylor, and R. M. Arnold (1992). Absolutely relative: How research results are summarized can affect treatment decisions. *The American Journal of Medicine* **92**(2): 121–124. DOI: 10.1016/0002-9343(92)90100-P.

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**(1): 3–39. DOI: 10.2307/1270359.

Gardner, W., E. P. Mulvey, and E. C. Shaw (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin* **118**(3): 392–404. DOI: 10.1037/0033-2909.118.3.392.

Glynn, R. J., T. A. Stukel, S. M. Sharp, T. A. Bubolz, J. L. Freeman, and E. S. Fisher (1993). Estimating the variance of standardized rates of recurrent events, with application to hospitalizations among the elderly in New England. *American Journal of Epidemiology* **137**(7): 776–786.

Goetz, J. N., R. H. Guthrie, and A. Brenning (2011). Integrating physical and empirical landslide susceptibility models using generalized additive models. *Geomorphology* **129**(3,4): 376–386. DOI: 10.1016/j.geomorph.2011.03.001.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52**(3): 681–700. DOI: 10.2307/1913471.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **52**(3): 443–452.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. 1st ed. London: Chapman & Hall.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters* **99**(3): 585–590. DOI: 10.1016/j.econlet.2007.10.015.

Greenland, S. (1996). Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology* **7**(5): 498–501. DOI: 10.1097/00001648-199609000-00008.

Greenland, S. (2004). Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology* **160**(4): 301–305. DOI: 10.1093/aje/kwh221.

Greenland, S. and P. Holland (1991). Estimating standardized risk differences from odds ratios. *Biometrics* **47**(1): 319–322. DOI: 10.2307/2532517.

Greenland, S. and J. M. Robins (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **41**(1): 55–68. DOI: 10.2307/2530643.

Greenwood, B. (2005). Interpreting vaccine efficacy. *Clinical Infectious Diseases* **40**(10): 1519–1520. DOI: 10.1086/429833.

Greenwood, M. and G. U. Yule (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* **83**(2): 255–279. DOI: 10.2307/2341080.

Grotvedt, L., H. Stigum, R. Hovengen, and S. Graff-Iversen (2008). Social differences in smoking and snuff use among Norwegian adolescents: A population based survey. *BMC Public Health* **8**: 322. DOI: 10.1186/1471-2458-8-322.

Gujarati, D. N. (2003). *Basic Econometrics*. 4th ed. New York: McGraw-Hill: p. 582.

Guo, J. Q. and P. K. Trivedi (2002). Flexible parametric models for long-tailed patent count distributions. *Oxford Bulletin of Economics and Statistics* **64**(1): 63–82. DOI: 10.1111/1468-0084.00004.

Hastie, T. J. (1992). "Generalized additive models". *Statistical Models in S*. ed. by J. M. Chambers and T. J. Hastie. Pacific Grove, CA: Wadsworth & Brooks/Cole. Chap. 7: pp. 249–307.

Hastie, T. J. (2013). *gam: Generalized Additive Models*. R package version 1.09. URL: http://CRAN.R-project.org/package=gam.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. 1st ed. London: Chapman & Hall.

Hauck, W. W., S. Anderson, and S. M. Marcus (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials* **19**(3): 249–256. DOI: 10.1016/S0197-2456(97)00147-5.

Hausman, J., B. H. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents-R & D relationship. *Econometrica* **52**(4): 909–938. DOI: 10.2307/1911191.

Hilbe, J. M. (2011). *Negative Binomial Regression*. 2nd ed. Cambridge: Cambridge University Press.

Hilbe, J. M. (2014). *COUNT: Functions, data and code for count data*. R package version 1.3.2. URL: http://CRAN.R-project.org/package=COUNT.

Hinde, J. (1982). "Compound Poisson regression models". *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*. **14**. New York: Springer-Verlag. Chap. 11: pp. 109–121.

Hiyoshi, A., Y. Fukuda, M. J. Shipley, M. Bartley, and E. J. Brunner (2013). A new theory-based social classification in Japan and its validation using historically collected information. *Social Science & Medicine* **87**: 84–92. DOI: 10.1016/j.socscimed.2013.03.021.

Holcomb Jr., W. L., T. Chaiworapongsa, D. A. Luke, and K. D. Burgdorf (2001). An odd measure of risk: Use and misuse of the odds ratio. *Obstetrics and Gynecology* **98**(4): 685–688. DOI: 10.1016/S0029-7844(01)01488-0.

Holgate, P. (1970). The modality of some compound Poisson distributions. *Biometrika* **57**(3): 666–667. DOI: 10.2307/2334787.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(2): 271–293. DOI: 10.1111/1467-9868.00125.

Joe, H. and R. Zhu (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal* **47**(2): 219–229. DOI: 10.1002/bimj.200410102.

Johnson, N. L., S. Kotz, and A. W. Kemp (1993). *Univariate Discrete Distributions*. 2nd ed. New York: Wiley-Interscience: pp. 199–235.

Karlis, D. (2001). A general EM approach for maximum likelihood estimation in mixed Poisson regression models. *Statistical Modelling* **1**(4): 305–318.

Keech, A. C., P. Mitchell, P. A. Summanen, J. O'Day, T. M. E. Davis, M. S. Moffitt, M. R. Taskinen, R. J. Simes, D. Tse, E. Williamson, A. Merrifield, L. T. Laatikainen, M. C. d'Emden, D. C. Crimet, R. L. O'Connell, and P. G. Colman (2007). Effect of fenofibrate on the need for laser treatment for diabetic retinopathy (FIELD study): A randomised controlled trial. *The Lancet* **370**(9600): 1687–1697. DOI: 10.1016/S0140-6736(07)61607-9.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science* **33**(3): 762–784. DOI: 10.2307/2111071.

Kleinbaum, D. G., L. L. Kupper, and H. Morgenstern (1982). *Epidemiologic Research: Principles and Quantitative Methods*. New York: Van Nostrand Reinhold: p. 359.

Koepsell, T. D. and N. S. Weiss (2003). *Epidemiologic Methods: Studying the Occurrence of Illness*. New York: Oxford University Press: pp. 196–199.

Kovalchik, S. A., R. Varadhan, B. Fetterman, N. E. Poitras, S. Wacholder, and H. A. Katki (2013). A general binomial regression model to estimate standardized risk differences from binary response data. *Statistics in Medicine* **32**(5): 808–821. DOI: 10.1002/sim.5553.

Kristiansen, I. S. and D. Gyrd-Hansen (2004). Cost-effectiveness analysis based on the number-needed-to-treat: common sense or non-sense? *Health Economics* **13**: 9–19. DOI: 10.1002/hec.797.

Lachin, J. M. (2011). *Biostatistical Methods: The Assessment of Relative Risk*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.: p. 14.

Lacy, C. R., J. A. Barone, D.-C. Suh, P. L. Malini, M. Bueno, D. M. Moylan, and J. B. Kostis (2001). Impact of presentation of research results on likelihood of prescribing medications to patients with left ventricular dysfunction. *The American Journal of Cardiology* **87**(2): 203–207. DOI: 10.1016/S0002-9149(00)01317-5.

Laupacis, A., D. L. Sackett, and R. S. Roberts (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* **318**(26): 1728–1733. DOI: 10.1056/NEJM198806303182605.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15**(3): 209–225. DOI: 10.2307/3314912.

Lee, J. (1981). Covariance adjustment of rates based on the multiple logistic regression model. *Journal of Chronic Diseases* **34**(8): 415–426. DOI: 10.1016/0021-9681(81)90040-0.

Leitenstorfer, F. and G. Tutz (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8**(3): 654–673. DOI: 10.1093/biostatistics/kxl036.

Lesnoff, M. and R. Lancelot (2012). *aod: Analysis of Overdispersed Data*. R package version 1.3. URL: http://CRAN.R-project.org/package=aod.

Lindsey, J. (2013). *gnlm: Generalized Nonlinear Regression Models*. R package version 1.0. URL: popgen.unimaas.nl/~jlindsey/rcode.html.

Liu, C. and D. B. Rubin (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**(4): 633–648. DOI: 10.2307/2337067.

Lord, D. and F. Mannering (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* **44**(5): 291–305. DOI: 10.1016/j.tra.2010.02.001.

Lord, D., S. P. Washington, and J. N. Ivan (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention* **37**(1): 35–46. DOI: 10.1016/j.aap.2004.02.004.

Lumley, T., R. Kronmal, and S. Ma (2006). *Relative risk regression in medical research: Models, contrasts, estimators, and algorithms*. Working Paper 293. UW Biostatistics Working Paper Series. URL: http://www.bepress.com/uwbiostat/paper293.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23**(19): 2937–2960. DOI: 10.1002/sim.1903.

Malloy, E. J., D. Spiegelman, and E. A. Eisen (2009). Comparing measures of model selection for penalized splines in Cox models. *Computational Statistics and Data Analysis* **53**(7): 2605–2616. DOI: 10.1016/j.csda.2008.12.008.

Marschner, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**(3): 666–683. DOI: 10.1198/jcgs.2010.09127.

Marschner, I. C. (2011). glm2: Fitting generalized linear models with convergence problems. *The R Journal* **3**(2): 12–15.

Marschner, I. C. (2014). Combinatorial EM algorithms. *Statistics and Computing* **24**(6): 921–940. DOI: 10.1007/s11222-013-9411-7.

Marschner, I. C. (2014). *glm2: Fitting Generalized Linear Models*. R package version 1.1.2. URL: http://CRAN.R-project.org/package=glm2.

Marschner, I. C. (2015). Relative risk regression for binary outcomes: Methods and recommendations. *Australian & New Zealand Journal of Statistics*. In press.

Marschner, I. C. and A. C. Gillett (2012). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics* **13**(1): 179–192. DOI: 10.1093/biostatistics/kxr030.

Marschner, I. C., A. C. Gillett, and R. L. O'Connell (2012). Stratified additive Poisson models: Computational methods and applications in clinical epidemiology. *Computational Statistics and Data Analysis* **56**(5): 1115–1130. DOI: 10.1016/j.csda.2011.08.002.

Maydeu-Olivares, A. (2005). "Linear item response theory, nonlinear item response theory, and factor analysis: A unified framework". *Contemporary Psychometrics*. Ed. by A. Maydeu-Olivares and J. J. McArdle. New Jersey: Lawrence Erlbaum Associates: pp. 73–100.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

McLachlan, G. J. (1997). On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research* **6**(1): 76–98.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons, Inc.

Mehrotra, D. V. and R. Railkar (2000). Minimum risk weights for comparing treatments in stratified binomial trials. *Statistics in Medicine* **19**(6): 811–825. DOI: 10.1002/(sici)1097-0258(20000330)19:6<811::aid-sim390>3.0.co;2-z.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**(2): 267–278. DOI: 10.2307/2337198.

Meng, X.-L. and D. Van Dyk (1997). The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(3): 511–567. DOI: 10.1111/1467-9868.00082.

Pearce, N. (2004). Effect measures in prevalence studies. *Environmental Health Perspectives* **112**(10): 1047–1050. DOI: 10.1289/ehp.6927.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**(4): 425–441. DOI: 10.1214/ss/1177012761.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **23**(2): 469–475.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**: 507–554. DOI: 10.1111/j.1467-9876.2005.00510.x.

Rothman, K. J. (1978). Occam's razor pares the choice among statistical models. *American Journal of Epidemiology* **108**(5): 347–349.

Rothman, K. J. and J. D. Boice (1982). *Epidemiologic Analysis with a Programmable Calculator*. 2nd ed. Chestnut Hill, MA: Epidemiology Resources Inc.: p. 12.

Rothman, K. J. and S. Greenland (1998). *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott–Raven: pp. 31–38.

Rothman, K. J., S. Greenland, and A. M. Walker (1980). Concepts of interaction. *American Journal of Epidemiology* **112**(4): 467–470.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**(4): 735–757. DOI: 10.1198/106186002853.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. New York: Cambridge University Press: p. 131.

Santos Silva, J. M. C. and F. Windmeijer (2001). Two-part multiple spell models for health care demand. *Journal of Econometrics* **104**(1): 67–89. DOI: 10.1016/S0304-4076(01)00059-8.

SAS Institute Inc. (2008). *SAS/STAT Software, Version 9.2*. Cary, NC. URL: http://www.sas.com/.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York: Springer-Verlag: p. 85.

Spiegelman, D. and E. Hertzmark (2005). Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* **162**(3): 199–200. DOI: 10.1093/aje/kwi188.

StataCorp (2013). *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.

Stijnen, T. and H. C. Van Houwelingen (1993). Relative risk, risk difference and rate difference models for sparse stratified data: A pseudo likelihood approach. *Statistics in Medicine* **12**(24): 2285–2303. DOI: 10.1002/sim.4780122406.

Stjernman, M., M. Green, Å. Lindström, O. Olsson, R. Ottvall, and H. G. Smith (2013). Habitat-specific bird trends and their effect on the Farmland Bird Index. *Ecological Indicators* **24**: 382–391. DOI: 10.1016/j.ecolind.2012.07.016.

Stukel, T. A., R. J. Glynn, E. S. Fisher, S. M. Sharp, G. Lu-Yao, and J. E. Wennberg (1994). Standardized rates of recurrent outcomes. *Statistics in Medicine* **13**(17): 1781–1791. DOI: 10.1002/sim.4780131709.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods* **7**(1): 13–26. DOI: 10.1080/03610927808827599.

Tang, W., H. He, and X. M. Tu (2012). *Applied Categorical and Count Data Analysis*. Boca Raton, FL: CRC Press: p. 182.

Thall, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data. *Biometrics* **44**(1): 197–209. DOI: 10.2307/2531907.

Thurston, S. W., M. P. Wand, and J. K. Wiencke (2000). Negative binomial additive models. *Biometrics* **56**(1): 139–144. DOI: 10.1111/j.0006-341X.2000.00139.x.

Tutz, G. and F. Leitenstorfer (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics* **16**(1): 165–188. DOI: 10.1198/106186007X180949.

Ukoumunne, O. C., E. Williamson, A. B. Forbes, M. C. Gulliford, and J. B. Carlin (2010). Confounder-adjusted estimates of the risk difference using propensity score-based weighting. *Statistics in Medicine* **29**(30): 3126–3136. DOI: 10.1002/Sim.3935.

Van de Werf, F., C. P. Cannon, A. Luyten, K. Houbracken, C. H. McCabe, S. Berioli, E. Bluhmki, H. Sarelin, F. Wang-Clow, N. L. Fox, and E. Braunwald (1999). Safety assessment of single-bolus administration of TNK tissue-plasminogen activator in acute myocardial infarction: The ASSENT-1 trial. *American Heart Journal* **137**(5): 786–791. DOI: 10.1016/S0002-8703(99)70400-X.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. 4th ed. New York: Springer: pp. 206–208.

Ver Hoef, J. M. and P. L. Boveng (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* **88**(11): 2766–2772. DOI: 10.1890/07-0043.1.

Wacholder, S. (1986). Binomial regression in GLIM: Estimating risk ratios and risk differences. *American Journal of Epidemiology* **123**(1): 174–184.

Walter, S. D. (2000). Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology* **53**(9): 931–939. DOI: 10.1016/S0895-4356(00)00210-9.

Walter, S. D. and T. R. Holford (1978). Additive, multiplicative, and other models for disease risks. *American Journal of Epidemiology* **108**(5): 341–346.

Wang, P., I. M. Cockburn, and M. L. Puterman (1998). Analysis of patent data: A mixed-Poisson-regression-model approach. *Journal of Business & Economic Statistics* **16**(1): 27–41. DOI: 10.2307/1392013.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**(3): 439–447. DOI: 10.2307/2334725.

Willan, A. R. and A. H. Briggs (2006). *Statistical Analysis of Cost-effectiveness Data*. Chichester, UK: John Wiley & Sons Ltd: p. 3.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. 5th ed. Berlin: Springer-Verlag.

Winkelmann, R. and K. F. Zimmermann (1991). A new approach for modeling economic count data. *Economics Letters* **37**(2): 139–143. DOI: 10.1016/0165-1765(91)90122-2.

Winkelmann, R. and K. F. Zimmermann (1992). "Robust Poisson regression". *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling, Munich, 13–17 July 1992*. New York: Springer-Verlag: pp. 201–206.

Wisniewski, T. K. M. (1966). Another statistical solution of a combinatorial problem. *The American Statistician* **20**(3): 25. DOI: 10.1080/00031305.1966.10480403.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(1): 3–36. DOI: 10.1111/j.1467-9868.2010.00749.x.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**(1): 95–103. DOI: 10.1214/aos/1176346060.

Xu, Y., Y. B. Cheung, K. F. Lam, S. H. Tan, and P. Milligan (2010). A simple approach to the estimation of incidence rate difference. *American Journal of Epidemiology* **172**(3): 334–343. DOI: 10.1093/aje/kwq099.

Yao, F. and T. C. M. Lee (2008). On knot placement for penalized spline regression. *Journal of the Korean Statistical Society* **37**(3): 259–267. DOI: 10.1016/j.jkss.2008.01.003.

Yeeles, A. (2015). Weathering unrest: The ecology of urban social disturbances in Africa and Asia. *Journal of Peace Research* **52**(2): 158–170. DOI: 10.1177/0022343314557508.