The estimation of Semiparametric Generalized Linear Models

By

Busayasachee Puang-Ngern

A thesis submitted to Macquarie University

for the degree of Doctor of Philosophy

Department of Mathematics and Statistics

Faculty of Science and Engineering

December 2018

MACQUARIE
University
SYDNEY·AUSTRALIA

Typeset in LaTeX 2_{ε}.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Busayasachee Puang-Ngern

Acknowledgements

Deepest appreciation and thanks to my supervisors, Associate Professor Jun Ma, for your always kindly support. Your guidance and teaching are valuable to me. To my co-supervisors, Associate Professor Ayse Bilgin and Dr Timothy Kyng, I really appreciate for your always being kindness and support. To all of my perfect supervisors, I am so grateful for all of your time and advice given to me along my journey.

I would like to thank Dr Thomas Fung for his guidance about PIT.

I am thankful to Chulalongkorn University in Thailand and Macquarie University in Australia for providing me the scholarship to pursue study in Australia.

Special thanks to Andrew Locke, Gustavo Thomas and P'Noi for their help on English proof-reading.

Thanks to the professional staff (MQU,OEA,CU) that always willing to help me to go through my paper work.

Thanks to all professors in the Department of Statistics for their friendliness.

Thanks to my friends (HDR students in the Department of Mathematics and Statistics, Thai friends, MQ Thai and International friends, P'Kwang, P'Paew, Housemate) to be my

good friends and make me enjoy life.

Finally, my warmest thanks to my family, my mom, my dad, my sister and Stubby for your patience and support.

Abstract

In this thesis, a novel method for fitting the semiparametric generalized linear model (SP-GLM) is developed and tested. We demonstrate that this provides an effective model fitting algorithm to the SP-GLM, particularly, when dealing with very large data sets. We also propose another special SP-GLM and discuss how to fit this special model. This special SP-GLM assumes the canonical link function, which simplifies the algorithm to fit this model.

GLMs are widely used for data analysis. However, in some applications, GLMs do not perform well in model fitting when the selected distribution for the response data is inaccurate. The SP-GLM with a nonparametric reference density extends the conventional GLMs. The SP-GLM offers flexibility in regression modelling by relaxing the requirement of a known response distribution in GLMs to only require that the response variable has a distribution from some exponential family. However, a limitation has been observed in the application of the existing SP-GLM method (Huang, 2014) on large data sets, presumably due to the significant increase in the number of constraints for the SP-GLM for large sample sizes. The proposed new SP-GLM methods in this thesis will enable to fit SP-GLM to very large data

sets.

In this research, the focus is on the regression coefficients estimations and inferences. An iterative algorithm is developed for estimation of the regression coefficients and the reference density simultaneously. The asymptotic properties of the estimators subject to active constraints are also provided.

Performance of the proposed methods are tested through simulation studies and real data applications. The simulation results have indicated effectiveness for the methods proposed in this research, with accurate estimation of the regression coefficients, as well as inference. The conclusion reached in this research is that the proposed model fitting methods enhance the capacity of the SP-GLM to handle very large data sets with fast convergence.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background and Aims

Generalized linear models (GLMs) (McCullagh & Nelder, 1983, 1989; Nelder & Wedderburn, 1972) are useful regression analysis tools. They extend the conventional linear regression modeling beyond requiring a normal distribution of the response variable to various other distributions and they also provide flexibility by allowing different link functions. A link

function is a function that makes the mean of a response variable share a linear relation with predictors. The well-known logistic regression is an example of a GLM.

GLMs have been employed in many areas such as actuarial science and medical research. In medical research, for example, GLMs are used to model healthcare costs and resource use (see e.g. Blough, Madden, and Hornbrook (1999); Manning and Mullahy (2001); Mihaylova, Briggs, O'hagan, and Thompson (2011)). The healthcare costs can be modeled using the Gamma distribution and the resource use can be modeled using the Poisson or negative binomial distribution. For actuarial science, the review of the utilization of GLMs in various problems including claims reserving, premium rating, multiple-state models, mortality and lapse rates can be found in Haberman and Renshaw (1996). GLMs have been used in actuarial science since the early 1980s (Haberman & Renshaw, 1996), for example, to support critical decisions made by insurance companies (De Jong & Heller, 2008) about life insurance, non-life insurance and pensions. In non-life insurance, GLMs can be applied to investigate the relationship between the annual claim frequency and the risk factors in vehicle insurance (e.g. Kafková, Křivánková, et al. (2014)). The number of claims can be modeled using the Poisson distribution with the log link. For life insurance, Renshaw and Haberman (1986) modeled the policy lapse rate using the binomial distribution with the logit link. Further examples of GLMs in actuarial science such as insurance cost pricing can be found in e.g. Cizek, Härdle, and Weron (2005); De Jong and Heller (2008); Denuit, Maréchal, Pitrebois, and Walhin (2007); Frees, Derrig, and Meyers (2014); Ohlsson and Johansson (2010). In this thesis, the applications to a health data set and an insurance data set are used to demonstrate the performance of our proposed methods.

Despite their broad usefulness in regression analysis in various fields as mentioned above, GLMs may not provide good solutions in some situations. In practice, one can easily find

examples where the response data may not be well modeled by the handful of parametric exponential family of distributions commonly used in GLMs.

GLMs are fully parametric in both the response distribution and the systematic component (see Section 2.1.2). For a GLM, two important components are the mean as a function of the linear predictors and the (conditional) distribution of the response variable. Both of these components must be fully defined. But in practice, sometimes one or both components cannot be specified satisfactorily. Many proposed extensions of GLMs attempt to relax these requirements by including a nonparametric component in either the mean function or the response distribution, or both. Most GLM extensions deal with a nonparametric component in the mean function. For example, the linear predictor in a generalized additive model (GAM) (Hastie & Tibshirani, 1990) depends linearly on unknown smooth functions. A single index nonparametric component in a generalized semiparametric single-index mixed model (GSSIMM) (Chen, 2010) is incorporated into the mean function. However, this type of GLM extension is not the focus of this thesis.

Our focus in this thesis is on the other way to extend GLMs. That is to include a nonparametric component into the response distribution so that the model becomes more flexible so as to accommodate departures from any conventional distribution chosen from the exponential family.

There are some statistical methods for relaxing the response distribution assumption. One popular method is the Quasi-likelihood method (Wedderburn, 1974) which does not require the response distribution at all. The Quasi-likelihood method only requires a relationship between the mean and the variance. This method is often described as robust in the sense that it can provide consistent regression coefficient estimates provided that the mean model is correctly specified (Crowder, 1986). However, as the actual probability distribution is not

considered, there is no further information about the probability mechanism which generated the data. Thus, it cannot be easily used to make inferences about the response distribution.

For the semiparametric GLM (SP-GLM) considered in this thesis, the distribution of the response variable is partly specified. This method is very useful for handling real world data, especially when there is limited prior knowledge of the response distribution. In fact, it assumes the distribution of the response variable contains both parametric and nonparametric components. This model can handle not only the standard distributions in the GLM framework, it can also handle some nonstandard distributions such as the distributions of zero-inflated and overdispersed count data. This relaxation on response distribution specification can avoid inefficient parameter estimates and inference due to model misspecification (e.g. Gardner, Mulvey, and Shaw (1995), White (1982), Drake (1993)). Even with an unspecified density function (the nonparametric component), this SP-GLM is in a full probability framework. This means we can gain more information for the probability mechanism that generates the data. This is useful for model selection and diagnostics as well as for making predictive inferences.

In this SP-GLM, the response density function is written in the form of an exponential tilting of a reference density. We assume the reference density f_0 is unknown and thus is an infinite dimensional parameter. It can be estimated simultaneously with the regression coefficient parameters. However, the fact of infinite dimension of the reference density makes it difficult to fit the model. A method to overcome this problem is by approximating the infinite dimensional parameter with a finite dimensional parameter.

The SP-GLM that applies exponential tilting to the response distribution was introduced by Rathouz and Gao (2009). Some favourable results for some count responses were shown in that paper. A computational algorithm has been developed for the response data with finite

and known support $C = [c_1, \ldots, c_U]^T$ with cardinality U. The reference density then becomes a discrete distribution $\tilde{\boldsymbol{f}}_{\boldsymbol{0}} = [f_0(c_1), \ldots, f_0(c_U)]^T$. In the estimation of $\tilde{\boldsymbol{f}}_{\boldsymbol{0}}$, $(U-2)$ components were first arbitrarily selected, denoted as $\boldsymbol{f}_{\boldsymbol{0}}^{*} = [f_0^*(c_1), \ldots, f_0^*(c_{U-2})]^T$. The remaining two parameters of $\tilde{\boldsymbol{f}}_{\boldsymbol{0}}$ were expressed as functions of $\boldsymbol{f}_{\boldsymbol{0}}^{*}$ to enforce two constraints on $\tilde{\boldsymbol{f}}_{\boldsymbol{0}}$, discussed in Section 2.3.3. The problem then becomes estimation of $\boldsymbol{h}_{\boldsymbol{0}}^{*}$ where $h_0^*(c_u) = \log\{f_0^*(c_u)\}$ for $u = 1, \ldots, U-2$. However, reparameterization of $\boldsymbol{f}_{\boldsymbol{0}}^{*}$ by logarithm may create local maxima, which means the starting value of an algorithm may determine the final solution.

Huang (2014) replaced the reference density by the probability masses $\boldsymbol{p} = [p_1, \ldots, p_n]^T$ supported on the observed response $[y_1, \ldots, y_n]^T$. This was motivated by the empirical likelihood approach of Owen (2001). For Huang's method, the constrained optimization solver function within the MATLAB optimization toolbox is used for parameter estimation. This computational method works well with small and medium size data sets (sample sizes of less than 800 observations). For the SP-GLM, the required number of constraints is at least as big as the number of observations in the data set. Using a built-in solver limits the sample sizes allowed for the model.

The concept and theory of SP-GLM is remarkable. However, the SP-GLM is not extensively used in practice. We aim to develop a novel method to fit the SP-GLM with an unspecified reference density. We develop our own computational algorithm, avoiding using any solver in MATLAB or R. The new method has an enhanced computational capability and is able to handle large data sets. In addition, we propose a SP-GLM with the canonical link function. In this model, the form of link function is not explicitly specified. This model is easier to fit than the SP-GLM with a user specified link function.

We just found recently that Aeberhard and Hannay (2018) and Wurm and Rathouz (2018) were independently working on the SP-GLM problem. The algorithms developed in this

thesis are different from theirs. In particular, we differ in the methods to estimate f_0 and the way to impose constraints on f_0. Aeberhard and Hannay (2018) approximates f_0 using the linear splines and then estimates its log transformation. They are currently exploring the constrained optimizers for estimating f_0. Wurm and Rathouz (2018) approximate f_0 using the same approximation as in Rathouz and Gao (2009), then the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), which is an iterative method to solve unconstrained optimization problems, is applied to obtain the estimate of log transformation of $\tilde{\boldsymbol{f}}_{\mathbf{0}}$.

1.2 Thesis outline

The rest of thesis is set out as follows. Chapter 2 provides backgrounds for GLM and SP-GLM, and it also summarizes the optimization methods needed in this thesis. Chapter 3 introduces the proposed new model fitting algorithm for the SP-GLM. The asymptotic properties of the parameters subject to the active constraints are derived. Various simulation results are presented to examine the performance of our proposed method. In Chapter 4, a special case of the SP-GLM is presented, where the canonical link function is used. Simulation studies are conducted to investigate the performance of this method. The applications of our model fitting methods to the real data sets are reported in Chapter 5. Then Chapter 6 provides conclusions and suggestions for future research. MATLAB codes for both methods are given in Appendix A.

2

Literature Review

In this chapter, we provide the background for statistical models and methods used to develop our model fitting algorithm for the SP-GLM. GLMs are fundamental models for understanding the SP-GLM, therefore we start with the background of GLMs in Section 2.1. Section 2.2 explains briefly the exponential tilting approach that is used to form the response distribution. Section 2.3 explains the SP-GLM which adopts exponential tilting and presents the model's identifiability issue. Existing methods for fitting SP-GLMs are discussed in this section.

Their computational issues and their ability to handle data problems are also considered. The maximum likelihood estimation is briefly explained in Section 2.4. Some useful constrained optimization methods, that are used to develop our model's algorithm, are illustrated in Section 2.5. Our proposed methods are briefly introduced in Section 2.6.

2.1 The generalized linear model

GLMs (McCullagh & Nelder, 1989) are an extension of linear models. They allow the distribution of the response variable to depart from the normal distribution. One benefit of GLMs is in their interpretation. The relationship between the expected response mean and the covariates can be interpreted through the link function. Due to these features, GLMs have become a popular and extensively used tool for data analysis.

We set the framework for the notation used throughout the thesis here. Suppose we have n independent response random variables, $Y_1, \ldots, Y_n$. The corresponding observations are $y_1, \ldots, y_n$. Let $y = [y_1 \ldots y_n]^T$ be the $n \times 1$ observed response vector. The design matrix $\mathbf{X} = \left[\mathbf{X}_1^T \ldots \mathbf{X}_n^T\right]^T$ is the $n \times q$ covariates matrix where q is the number of covariates and $\mathbf{X}_i = [x_{i1} \ldots x_{iq}]$ is the row vector of $\mathbf{X}$ for the ith observation where $i = 1, ..., n$. For simplicity, conditioning on the covariates is implicit throughout.

GLMs are comprised of three components: a probability distribution, a linear predictor, and a link function (see also, for example, Dobson and Barnett (2008); Fahrmeir and Tutz (2013); De Jong and Heller (2008)).

2.1.1 The probability distribution

The first component of a GLM is a (conditional) probability distribution for the response variable. This is sometimes also known as the error model. In linear regression, the error

model is restricted to be a normal distribution. However, in GLMs, the probability distribution of Y_i can be any distribution that is a member of an exponential family of distributions. Thus the probability density function of Y_i can be expressed in the form

$$f(y_i) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i)\}, \tag{2.1}$$

where θ_i is an unknown canonical parameter, $b(\theta_i)$ is known and relates to the normalizing constant of the distribution and $c(\cdot)$ is a known function. θ_i is tied to the mean and thus is linked to the linear predictor. This is a rich family of distributions but in practice GLMs use only a few of its members. Well-known exponential family distributions include, for example, the normal, Poisson, binomial and gamma distributions. Note that the density function of the exponential family can include a dispersion parameter ϕ. This is called the exponential dispersion family. The density function is then in the form:

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{2.2}$$

where $a(\cdot)$ is a known function. In general $a_i(\phi) = \frac{\phi}{w_i}$ where w_i is the known weight of observations which is typically set to 1.

2.1.2 The linear predictor

The second component of a GLM is the linear predictor. This is a systematic component. It is specified as

$$\eta_i = \mathbf{X}_i\boldsymbol{\beta}, \tag{2.3}$$

where $\boldsymbol{\beta}$ is a q-column vector of the regression coefficient parameters. This $\mathbf{X}_i\boldsymbol{\beta}$ is a linear combination of covariates similar to the linear regression setting.

2.1.3 The link function

The third component of a GLM is the link function $g(\cdot)$ which is user specified and is required to be monotonic. It is used to link the response mean and the linear predictor together. That is

$$g(\mu_i) = \eta_i, \tag{2.4}$$

where μ_i is the conditional mean of Y_i given the covariates. The advantage of the link function is that the non-linear structure can be transformed to have a linear structure through the link function. This feature generalizes the traditional linear relationship of the response mean and the regression coefficient parameters. The mean function is denoted by

$$\mathrm{E}(Y_i|\mathbf{X}_i) = \mu_i(\mathbf{X}_i;\boldsymbol{\beta}) \equiv \mu_i = g^{-1}(\eta_i), \tag{2.5}$$

where $g^{-1}(\cdot)$ is the inverse of the link function. The mean function is a smooth and invertible function of the linear predictor.

Some typically used link functions are shown in Table 2.1. These functions can also be used in the SP-GLM.

Table 2.1: Commonly used link functions

Link function	$g(\mu_i)$	$g^{-1}(\eta_i)$	$\frac{\partial \mu_i}{\partial \eta_i}$
identity	μ_i	η_i	1
log	$\ln(\mu_i)$	$\exp(\eta_i)$	$\exp(\eta_i)$
logit	$\ln\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$	$\frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2}$
inverse	μ_i^{-1}	η_i^{-1}	$-\eta_i^{-2}$

The canonical link function

If the $g(\cdot)$ function is chosen such that $\theta_i = g(\mu_i)$, then $g(\cdot)$ is called the canonical link function. Examples of canonical links for well-known exponential family distributions are the log link for the Poisson distribution, logit link for the binomial distribution, identity link for the normal distribution and inverse link for the gamma distribution. Although the canonical link function is a useful link function, sometimes a non-canonical link function may be preferable, because it may provide better model fitting for some data sets or it may be easier to interpret the relationship of the variables, such as the log link is more likely to be used with the gamma and negative binomial distribution than their canonical links. However, there are some computational difficulties for a non-canonical link in comparison to a canonical link. For example, using the gamma distribution with the identity link involves numerical issues since the identity link is not guaranteed to produce positive expectation as required for the gamma distribution.

Generally, the canonical link is used as a default link function due to its desirable mathematical and statistical properties. It generally guarantees the mean constraint to be within the range of the response variable (Breheny, 2013). For example, the canonical link for the Poisson distribution is the log link. That is $\log(\mu_i) = \eta_i = \theta_i$, then $\mu_i = \exp\{\theta_i\} \in [0, \infty]$ which is within the boundaries of the Poisson distribution. Another example is the logit link which is the canonical link for the binomial distribution. The response mean is $\mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} \in [0, 1]$ which assures that mean lies within its boundaries.

Using the canonical link also simplifies the derivation of the maximum likelihood estimator (Breheny, 2013). Many linear regression properties such as sum of the residuals which are differences between observed and predicted responses, equals zero still hold true (Breheny, 2013) under the canonical link. The minimal sufficient statistic for $\boldsymbol{\beta}$ also exists

(Lindsey, 1997; Rodriguez, 2007). We use the canonical link function later in Chapter 4 for the special case of the SP-GLM.

2.1.4 The mean and variance functions

For a distribution from the exponential family, the relationship between the mean and the canonical parameter θ_i is specified as $\mu_i = b'(\theta_i)$ and the relationship between the variance and θ_i is specified as $\text{Var}(Y_i|\mathbf{X}_i) = a_i(\phi)b''(\theta_i)$.

The closed forms of θ_i, $b(\theta_i)$, $b'(\theta_i)$, and $b''(\theta_i)$ can be defined from the known response distribution. For example, let Y_i follow a Poisson distribution with probability mass function

$$f(y_i) = \frac{\mu_i^{y_i} \exp\{-\mu_i\}}{y_i!}.$$

This $f(y_i)$ can be expressed in the form of the exponential family,

$$f(y_i) = \exp\left\{y_i \log(\mu_i) - \mu_i - \log(y_i!)\right\},$$

where $\theta_i = \log(\mu_i)$, $b(\theta_i) = \exp\{\theta_i\}$, $\phi = 1$, $w_i = 1$ and $c(y_i, \phi) = -\log(y_i!)$. So the mean and variance for the Poisson distribution are $b'(\theta_i) = \exp\{\theta_i\} = \mu_i$ and $\text{Var}(Y_i|\mathbf{X}_i) = \frac{\phi}{w_i} b''(\theta_i) = \exp\{\theta_i\} = \mu_i$.

Another example we wish to discuss here is the normal distribution (or Gaussian distribution) with mean μ_i and variance σ^2. The probability density function is

$$\begin{aligned} f(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}, \end{aligned}$$

where $\theta_i = \mu_i$, $b(\theta_i) = \frac{\theta_i^2}{2}$, $\phi = \sigma^2$, $w_i = 1$ and $c(y_i, \phi) = -\frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right)$. Then the mean is $b'(\theta_i) = \theta_i = \mu_i$ and the variance is $\text{Var}(Y_i|\mathbf{X}_i) = \frac{\phi}{w_i} b''(\theta_i) = \sigma^2$.

More examples of exponential family distributions, such as the binomial, gamma, and inverse Gaussian distributions, can be found in many GLMs books (for example, De Jong and Heller (2008)).

In many cases in applied analysis, the assumption of a known response distribution may be an impossible requirement. We can relax that requirement by assuming a particular structure for the response distribution that includes a nonparametric component.

2.2 Exponential tilting

We will first give some background on exponential tilting because it is a technique that can be used to build the response distribution in the SP-GLM as explained in Section 2.3.

Exponential tilting is a distribution shifting technique commonly used in simulation. It was proposed by Esscher (1932), then developed by Daniels (1954), and Barndorff-Nielsen and Cox (1979). Exponential tilting is also known as the Esscher transform in mathematical finance (Asmussen & Glynn, 2007, p. 330) and insurance (Cruz, Peters, & Shevchenko, 2015, p. 784).

For simplicity, consider a scalar random variable Y that has a probability distribution F_0 with density function f_0 with respect to some measure and moment generating function $M(\theta) = \mathrm{E}[\exp\{y\theta\}] < \infty$. Define $I\{Y \in \mathrm{d}y\}$ to be an indicator function: $I\{Y \in \mathrm{d}y\} = 1$ for $Y \in \mathrm{d}y$ and 0 otherwise, where $\mathrm{d}y$ is some infinitesimally small measurable set. The tilted distribution F can be generated from the original distribution F_0. The exponentially tilted distribution is in the form,

$$F(Y \in \mathrm{d}y) = \frac{\mathrm{E}\left[\exp\{y\theta\} I\{Y \in \mathrm{d}y\}\right]}{M(\theta)} = \exp\left\{y\theta - \kappa(\theta)\right\} F_0(Y \in \mathrm{d}y),$$

where $\kappa(\theta) = \log\left(\mathrm{E}\left[\exp\{y\theta\}\right]\right)$ is the cumulant generating function.

The exponentially tilted distribution function F has a density function that can be expressed as

$$f(y) = \frac{\exp\{y\theta\} f_0(y)}{M(\theta)} = \exp\{y\theta - \kappa(\theta)\} f_0(y). \tag{2.6}$$

This density function seems to be very similar in form to an exponential family distribution (2.1). But they are slightly different in that $f_0(y)$ in (2.6) is a density, while $\exp\{c(y)\}$ in (2.1) may not be a density. In other words, exponential tilting can construct distributions of the exponential family where the probability distribution F_0 is tilted by θ.

2.3 The semiparametric generalized linear model

There are many types of SP-GLMs. Most SP-GLMs deal with a nonparametric mean function and a parametric response distribution such as the generalized additive model (Hastie & Tibshirani, 1990) and the generalized semiparametric single-index mixed model (Chen, 2010). However, throughout this thesis, we consider only the SP-GLM that was introduced by Rathouz and Gao (2009). In this scenario, the mean function is a parametric function, while the response distribution contains a nonparametric component.

2.3.1 Model

The SP-GLM consists of the same three components as for GLMs: a probability distribution, a linear predictor, and a link function. Suppose each response variable Y_i has an exponentially tilted distribution F_i with support $\mathcal{Y} \subseteq \mathcal{R}$. Thus for a given $\mathbf{X}_i$, observation y_i is independently sampled from F_i with probability density function

$$f_i(y|\mathbf{X}_i) = f_0(y) \exp\{y\theta_i - b(\theta_i; f_0)\}, \tag{2.7}$$

where the cumulant generating function is

$$b(\theta_i; f_0) = \log \int_{\mathcal{Y}} \exp\{y\theta_i\} f_0(y) \mathrm{d}y.$$

This specification has the advantage that the response distribution is not required to be completely known *a priori*.

The exponential tilting provides a special group in the exponential family where $a(\phi) = 1$ and $\exp\{c(y, \phi)\}$ is a density $f_0(y)$. In GLMs, $f_0(y)$ is fully known from the pre-specified response distribution. In contrast, in this SP-GLM, $f_0(y)$ is unknown and can be estimated from the data. Therefore f_0 is nonparametric and thus is infinite dimensional and this makes the task of model fitting more difficult than for the standard GLMs.

The SP-GLM can cover a rich family of distributions by having different f_0. Flexibility of f_0 will enable better fits of density to data. For example, f_0 may have inflated probability mass at zero for zero-inflated data; f_0 may have thicker tails than the Poisson distribution for overdispersed count data.

As f_0 is seen as a density function, the required constraints on f_0 are: (i) $f_0(y) \geq 0$ (ii) $\int_{\mathcal{Y}} f_0(y)\mathrm{d}y = 1$, which are the properties of a density function.

The linear predictor, the link function and the mean function are specified in the same way as in the classical GLMs in (2.3), (2.4) and (2.5) respectively. The SP-GLM also has the same mean and variance formulae as GLM:

$$\mu_i = b'(\theta_i; f_0) = \int_{\mathcal{Y}} y \exp\left\{y\theta_i - b(\theta_i; f_0)\right\} f_0(y)\mathrm{d}y,$$

$$\mathrm{Var}(Y_i|\mathbf{X}_i) = b''(\theta_i; f_0) = \int_{\mathcal{Y}} (y - \mu_i)^2 \exp\left\{y\theta_i - b(\theta_i; f_0)\right\} f_0(y)\mathrm{d}y.$$

However, unlike in GLMs where the canonical parameter θ_i is known from the distribution specification, in the SP-GLM θ_i is unknown. It can be computed from the following

relationship:

$$g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) = \int_{\mathcal{Y}} y \exp\{y\theta_i - b(\theta_i; f_0)\} f_0(y)\mathrm{d}y.$$

Thus θ_i can be seen as a function of $\boldsymbol{\beta}$ and f_0.

2.3.2 Identifiability

Note that f_0 in the probability density function (2.7) is not identifiable. This is because we can get the same density $f_i(y|\mathbf{X}_i)$ by replacing f_0 with an exponentially tilted f_0. More specifically, if letting the exponential tilting of f_0 be

$$f_0^*(y) = f_0(y)\exp\{y\theta_0^* - b(\theta_0^*; f_0)\}, \tag{2.8}$$

where θ_0^* can be any finite constant value and

$$b(\theta_0^*; f_0) = \log \int_{\mathcal{Y}} \exp\left\{y\theta_0^*\right\} f_0(y)\mathrm{d}y. \tag{2.9}$$

Then we can demonstrate that $f_i(y|\mathbf{X}_i)$ can be expressed as

$$f_i(y|\mathbf{X}_i) = f_0^*(y)\exp\{y\theta_i^* - b^*(\theta_i^*; f_0^*)\}, \tag{2.10}$$

where

$$b^*(\theta_i^*; f_0^*) = \log \int_{\mathcal{Y}} \exp\left\{y\theta_i^*\right\} f_0^*(y)\mathrm{d}y. \tag{2.11}$$

Substituting $f_0^*(y)$ from (2.8) in (2.10), we get

$$f_i(y|\mathbf{X}_i) = f_0(y)\exp\left\{y(\theta_0^* + \theta_i^*) - \left(b(\theta_0^*; f_0) + b^*(\theta_i^*; f_0^*)\right)\right\}. \tag{2.12}$$

If we let $\theta_i^* = \theta_i - \theta_0^*$, we get $b^*(\theta_i^*; f_0^*) = b(\theta_i; f_0) - b(\theta_0^*; f_0)$ after substituting $f_0^*(y)$ from (2.8) into (2.11). Then (2.12) is the same as the conditional density in (2.7).

Note that $f_0^*(y)$ becomes unique if its mean value μ_0^*, i.e. $\int_{\mathcal{Y}} y f_0^*(y)\mathrm{d}y = \mu_0^*$, is fixed (Rathouz & Gao, 2009). The density function $f_i(y|\mathbf{X}_i)$ is invariant to the value of μ_0^* as

long as it is chosen from the interior of $\mathcal{Y}$ (Rathouz & Gao, 2009). One way to impose this identifiability constraint is by pre-specifying μ_0^* and solving for the corresponding θ_0^*, then for a given $f_0(y)$, $f_0^*(y)$ can be obtained (Wurm & Rathouz, 2018).

In Chapter 3, to make the model identifiable, we choose μ_0^* in the way such that $\theta_0^* = 0$. This is because if we force θ_0^* to be zero, then θ_i^* will be identical to θ_i and $b(\theta_0^*; f_0) = 0$, then $f_0^*(y) = f_0(y)$. Note that this special μ_0^* needs not to be specified explicitly.

2.3.3 Existing methods

Rathouz and Gao's method

Rathouz and Gao (2009) introduced the SP-GLM and provided parameter estimation algorithm for the response data with known finite support $C = [c_1, \ldots, c_U]^T$ with cardinality U. In their algorithm, the vector of probabilities is defined as $\tilde{\boldsymbol{f}}_{\boldsymbol{0}} = [f_0(c_1), \ldots, f_0(c_U)]^T$. The parameters $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{f}}_{\boldsymbol{0}}$ were estimated by using a maximum likelihood approach. The three constraints on $\tilde{\boldsymbol{f}}_{\boldsymbol{0}}$ are: (i) $0 \leq f_0(c_u) \leq 1$ for $u = 1, \ldots, U$; (ii) $\sum_{u=1}^{U} f_0(c_u) = 1$ and (iii) $\sum_{u=1}^{U} c_u f_0(c_u) = \mu_0^*$, where μ_0^* is a pre-selected quantity. The last constraint is the identifiability constraint which can be enforced differently (see Section 2.3.2). The maximum likelihood estimation of $\tilde{\boldsymbol{f}}_{\boldsymbol{0}}$ that satisfies these constraints is achieved by arbitrarily choosing $(U-2)$ elements from $\tilde{\boldsymbol{f}}_{\boldsymbol{0}}$ to form a new vector which is denoted as $\boldsymbol{f}_{\boldsymbol{0}}^{\boldsymbol{*}} = [f_0^*(c_1), \ldots, f_0^*(c_{U-2})]^T$. Then the remaining two elements $[f_0(c_{U-1}), f_0(c_U)]^T$ were expressed as functions of $\boldsymbol{f}_{\boldsymbol{0}}^{\boldsymbol{*}}$ through equations (ii) and (iii), and they can be obtained from $\hat{\boldsymbol{f}}_{\boldsymbol{0}}^{\boldsymbol{*}}$. Let a vector $\boldsymbol{h}_{\boldsymbol{0}}^{\boldsymbol{*}}$ be defined as $\boldsymbol{h}_{\boldsymbol{0}}^{\boldsymbol{*}} = [h_0^*(c_1), \ldots, h_0^*(c_{U-2})]^T$, the non-negativity constraints $f_0(c_u) \geq 0$ were imposed by reparameterizing $\boldsymbol{f}_{\boldsymbol{0}}^{\boldsymbol{*}}$ such that $f_0^*(c_u) = \exp\{h_0^*(c_u)\}$ for $u = 1, \ldots, U-2$. The Fisher scoring method can be applied to estimate $\boldsymbol{h}_{\boldsymbol{0}}^{\boldsymbol{*}}$, where a step size of, for example, $1/2$ can multiply the Fisher information matrix for $\boldsymbol{h}_{\boldsymbol{0}}^{\boldsymbol{*}}$ until the constraint $f_0(c_u) \leq 1$ is satisfied.

For given μ_i and $\tilde{\boldsymbol{f}}_{\mathbf{0}}$, θ_i can be solved from the mean condition $\mu_i = b'(\theta_i; \tilde{\boldsymbol{f}}_{\mathbf{0}})$. However, since $\mu_i \in (m, M)$ where $m = \inf(C)$ and $M = \sup(C)$, an ad-hoc method using the logit transform was applied to stabilize the solution. The logit transform function was defined as $g_l(\mu_i) = \text{logit}\left(\frac{\mu_i - m}{M - m}\right) = \log\left(\frac{\mu_i - m}{M - \mu_i}\right)$, and then the Newton's method is used to estimate θ_i which solves for θ_i from $g_l(\mu_i) = g_l(b'(\theta_i; \tilde{\boldsymbol{f}}_{\mathbf{0}}))$.

In the paper of Rathouz and Gao (2009), the numerical results of some polytomous responses for small cardinality U were shown.

Huang's method

Huang (2014) approximated f_0 by the non-negative empirical probability masses $\boldsymbol{p}$, which is a vector containing p_i's for $i = 1, \ldots, n$. $\boldsymbol{\beta}$ and $\boldsymbol{p}$ are estimated simultaneously via maximum empirical likelihood estimation. The empirical log-likelihood function is

$$l(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} (\log p_i + b_i + \theta_i y_i),$$

where $b_i = -b(\theta_i; \boldsymbol{p})$ are normalizing constants and both θ_i and b_i are treated as unknown parameters in the model fitting algorithm.

The MATLAB built-in optimization function named `fmincon()` is used to estimate β_j, $\log p_i$, b_i and θ_i for $j = 1, \ldots q$ and $i = 1, \ldots n$ that satisfies two constraints: (i) the normalization constraint

$$\sum_{i=1}^{n} p_u \exp\{b_i + \theta_i y_u\} = 1, \qquad \text{for} \quad u = 1, \ldots, n$$

and (ii) the mean constraint

$$\sum_{u=1}^{n} y_u p_u \exp\{b_i + \theta_i y_u\} = g^{-1}(\mathbf{X}_i \boldsymbol{\beta}), \qquad \text{for} \quad i = 1, \ldots, n.$$

The optimization scheme is set as follows:

1. Set the initial values of all parameters. Specifically, $\log p_i^{(0)} = \log(n^{-1})$, $b_i^{(0)} = 0$ and $\theta_i^{(0)} = 0$ for all i where n is the number of observations. $\beta_j^{(0)} = g(\bar{y})$ for $j = 1$ and for other j, $\beta_j^{(0)} = 0$, where $\bar{y}$ is the average of the observed response values.

2. At iteration k, set the options for the MATLAB optimization function as follows.

   (a) Set the algorithm to be the interior point algorithm.

   (b) Set the maximum number of iteration to be 10^5.

   (c) Set the convergence criterion of this optimization function to be

      i. $\max_j |\beta_j^{(k+1)} - \beta_j^{(k)}| < 10^{-8}$,

      ii. $\max_i |\log p_i^{(k+1)} - \log p_i^{(k)}| < 10^{-8}$,

      iii. $\max_i |b_i^{(k+1)} - b_i^{(k)}| < 10^{-8}$,

      iv. $\max_i |\theta_i^{(k+1)} - \theta_i^{(k)}| < 10^{-8}$ and

      v. $|l^{(k+1)} - l^{(k)}| < 10^{-8}$.

   (d) Set the normalization constraint to be $1 - \sum_{i=1}^{n} p_u^{(k)} \exp\{b_i^{(k)} + \theta_i^{(k)} y_u\} = 0$ for $u = 1, \ldots, n$.

   (e) Set the mean constraint to be $g^{-1}(\mathbf{X}_i \boldsymbol{\beta}^{(k)}) - \sum_{u=1}^{n} y_u p_u^{(k)} \exp\{b_i^{(k)} + \theta_i^{(k)} y_u\} = 0$ for all i.

3. In each iteration, the negative log-likelihood is computed according to $l^{(k)} = -\sum_{i=1}^{n} \left(\log p_i^{(k)} + b_i^{(k)} + \theta_i^{(k)} y_i \right)$. Note that the negative log-likelihood function is used since `fmincon()` searches for a minimum.

4. The stopping criteria is either when the maximum number of iteration is reached or the convergence criterion is satisfied.

This MATLAB function can only handle sample sizes of less than 800 observations (tested with a computer having Intel Core i5 CPU, 2.80 GHz, RAM 8 GB, 64-bit OS).

Wurm and Rathouz's method

As mentioned in Chapter 1, we recently found the paper (in press) of Wurm and Rathouz (2018) that proposed an alternative algorithm to fit the SP-GLM.

Recall that the observed support is $C = [c_1, \ldots, c_U]^T$. In general, $U = n$, but it is possible that $U < n$ if there are ties in the observed response values. In Wurm and Rathouz's algorithm, f_0 is approximated by $\tilde{\boldsymbol{f}}_{\mathbf{0}} = [f_0(c_1), \ldots, f_0(c_U)]^T$, the corresponding probability masses at $c_1, \ldots, c_U$. Note that this approximation method for f_0 is the same as Rathouz and Gao (2009).

An iterative approach is applied to find the optimal solutions of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{f}}_{\mathbf{0}}$ that maximizes the log-likelihood,

$$l(\boldsymbol{\beta}, \tilde{\boldsymbol{f}}_{\mathbf{0}}) = \sum_{i=1}^{n} \left(\theta_i y_i - \log \sum_{u=1}^{U} f_0(c_u) \exp\{\theta_i c_u\} + \sum_{u=1}^{U} I(y_i = c_u) \log f_0(c_u) \right),$$

subject to the constraints:

1. $g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) \in (m, M)$ for $i = 1, \ldots, n$ where $m \equiv \min\{c_1, \ldots, c_U\}$ and $M \equiv \max\{c_1, \ldots, c_U\}$;
2. $f_0(c_u) \geq 0$ for $u = 1, \ldots, U$;
3. $\sum_{u=1}^{U} f_0(c_u) = 1$ and
4. $\sum_{u=1}^{U} c_u f_0(c_u) = \mu_0^*$ where μ_0^* is some pre-specified value within the range (m, M).

$\tilde{\boldsymbol{f}}_{\mathbf{0}}$ is transformed using log function, i.e. $\tilde{\boldsymbol{g}}_{\mathbf{0}} = [g_0(c_1), \ldots, g_0(c_U)]^T = [\log f_0(c_1), \ldots, \log f_0(c_U)]^T$, to impose the non-negativity constraints on $\tilde{\boldsymbol{f}}_{\mathbf{0}}$. The estimate of $\tilde{\boldsymbol{g}}_{\mathbf{0}}$ is then achieved by solving the unconstrained optimization problems using the Broyden-Fletcher-

Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). To enforce increasing log-likelihood after updating $\tilde{\boldsymbol{g}}_{\mathbf{0}}$, the newly updated $\tilde{\boldsymbol{g}}_{\mathbf{0}}$ at iteration $k+1$ is half-step backtracked, i.e. continuously reupdates $\tilde{\boldsymbol{g}}_{\mathbf{0}}^{(k+1)} = \frac{1}{2}(\tilde{\boldsymbol{g}}_{\mathbf{0}}^{(k+1)} + \tilde{\boldsymbol{g}}_{\mathbf{0}}^{(k)})$ until the log-likelihood increases. After obtaining the estimate of $\tilde{\boldsymbol{g}}_{\mathbf{0}}$ at each BFGS iteration, the $\tilde{\boldsymbol{f}}_{\mathbf{0}}$ estimate can be obtained. According to the procedure described in Wurm and Rathouz (2018)'s paper, they then rescale the estimate $f_0(c_u)$ for all u, to the results satisfy the constraint $\sum_{u=1}^{U} f_0(c_u) = 1$. Then they solve for θ_0^* from equation $\sum_{u=1}^{U} c_u f_0(c_u) \exp\{\theta_0^* c_u\} / \sum_{u=1}^{U} f_0(c_u) \exp\{\theta_0^* c_u\} = \mu_0^*$. Finally, all estimates $f_0(c_u)$ are exponentially tilted to give $f_0(c_u) \exp\{\theta_0^* c_u\} / \sum_{v=1}^{U} f_0(y_v) \exp\{\theta_0^* y_v\}$. Note that this rescaling strategy may cause the solution to be suboptimal.

For the estimation of θ_i, Wurm and Rathouz (2018) use the same estimation procedure as Rathouz and Gao (2009) (see page 18), that can restrict $g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$ in its boundary (m, M). However, if $g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$ is close to a boundary, θ_i will converge to $\pm\infty$ which can cause numerical instability. To overcome this issue, a limit on the maximum value of $|\theta_i|$ could be placed (such as 500) as a default value.

Aeberhard and Hannay's method

Aeberhard and Hannay (2018) also provides an alternative model fitting method for the SP-GLM. As there is no published paper on Aeberhard and Hannay's method yet, the following information is based on Aeberhard and Hannay's presentation at the 2nd International Conference on Econometrics and Statistics on 21 June 2018. Aeberhard and Hannay (2018) uses a linear B-spline to approximate $\log f_0$. The data are viewed as if they are independent and identically distributed to obtain good starting values of f_0. However, a particular optimization method has not been set yet.

Note that the inverse of the information matrix for f_0 is required for the estimation of f_0 in Rathouz and Gao (2009)'s method. If this matrix has high dimension, then the computational costs are high. In addition, Rathouz and Gao (2009), Huang (2014), Wurm and Rathouz (2018) and Aeberhard and Hannay (2018) all adopted logarithm transformation of f_0 to impose the non-negativity constraints on f_0. The risk is this transformation may generate multiple local maxima. The main reason is that the second derivative matrix under the transformed parameter is no longer negative definite. Hence the final solution may depend on the starting value. Another problem is that this approach will exclude the possibility of $f_0 = 0$.

2.4 Maximum likelihood estimation

Our parameter estimation in Chapters 3 and 4 are based on the constrained maximum likelihood estimation. Thus in this section, we briefly explain the maximum likelihood estimation which is the basis for the constrained maximum likelihood estimation. Maximum likelihood estimation is a method used to estimate the unknown parameters based on the given sample. Maximum likelihood estimation chooses the parameter estimates that maximize the likelihood (or equivalently log-likelihood) function for a given sample.

In parametric GLMs, maximum likelihood estimation can be used to estimate $\boldsymbol{\beta}$. The probability function (2.2) for each Y_i is now denoted as $f(y_i; \boldsymbol{\beta})$ to emphasise its dependence on $\boldsymbol{\beta}$. All random variables Y_i are assumed independent. The likelihood function can be expressed as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\beta}),$$

and the log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\beta}).$$

Maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ are the solutions that maximize the log-likelihood. Often it is obtained by solving the following equation

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0.$$

The maximum likelihood estimator has many desirable properties such as functional invariance, i.e. if $g(\boldsymbol{\beta})$ is a monotonic transformation of $\boldsymbol{\beta}$ and the maximum likelihood estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$, then the maximum likelihood estimator of $g(\boldsymbol{\beta})$ is $g(\hat{\boldsymbol{\beta}})$. For a sufficiently large sample size, the maximum likelihood estimator has the minimum variance in the class of unbiased estimators and is consistent (i.e. $\hat{\boldsymbol{\beta}}$ converges in probability to its true value as $n \to \infty$). However, a closed form of the maximum likelihood estimator is not always available. In this situation, an iterative numerical method can be applied to find the maximum likelihood estimate. There are many different iterative methods and the user often selects an algorithm based on speed and reliability. Examples of famous optimization methods are Newton's method and Fisher scoring method. More details of these methods can be found in, for example, Thisted (1988).

2.5 Constrained optimization method

We propose a constrained optimization algorithm to estimate the regression coefficients and f_0. The SP-GLM described in Chapters 3 and 4 will have to deal with both the equality and the inequality constraints. Therefore in this section, some constrained optimization methods that can be used to impose the constraints for the SP-GLM are explained.

2.5.1 Lagrange Multipliers

The method of Lagrange multipliers is a convenient and well-known optimization method for imposing equality constraints. It can find the local maxima or minima of an objective function subject to equality constraints.

For simplicity, we explain the method of Lagrange multipliers for one parameter and a single equality constraint, although our methods in Chapters 3 and 4 have more than one constraints. Let β be a parameter of interest. Define $l(\beta)$ to be a log-likelihood function. Suppose the optimization problem is to maximize $l(\beta)$ subject to equality constraint $h(\beta) = 0$. l and h are assumed to be continuously differentiable. The Lagrangian is in the form

$$\mathscr{L}(\beta, \lambda) = l(\beta) - \lambda h(\beta),$$

where λ is a Lagrange multiplier. An optimal solution of the original constrained problem coincides with an unconstrained stationary point of the Lagrangian, but not always vice versa. The method of Lagrange multipliers produces a necessary condition for constrained optimization problems. The method of Lagrange multipliers is to solve for $\hat{\beta}$ that is a solution of

$$\frac{\partial \mathscr{L}(\beta, \lambda)}{\partial \beta} = 0 \quad \text{and} \quad \frac{\partial \mathscr{L}(\beta, \lambda)}{\partial \lambda} = 0.$$

In other words,

$$\frac{\partial l(\beta)}{\partial \beta} = \lambda \frac{\partial h(\beta)}{\partial \beta} \quad \text{and} \quad h(\beta) = 0.$$

2.5.2 KKT conditions

The Karush-Kuhn-Tucker (KKT) conditions (Kuhn & Tucker, 1951) generalize the method of Lagrange multipliers to handle equality and inequality constraints. Given that some regularity conditions are satisfied, the KKT conditions provide first-order necessary conditions for the

following optimal solution.

Suppose we want to find the solution to

$$\arg\max_{\beta} l(\beta)$$

subject to $h(\beta) = 0$ and $g(\beta) \geq 0$ where h and g are the equality and inequality constraint functions, respectively. Assume all l, g and h are continuously differentiable. The Lagrangian is

$$\mathscr{L}(\beta, \lambda, \nu) = l(\beta) - \lambda h(\beta) + \nu g(\beta),$$

where λ and ν are Lagrange multipliers. The KKT necessary conditions for the optimal solution are

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta} - \lambda \frac{\partial h(\beta)}{\partial \beta} + \nu \frac{\partial g(\beta)}{\partial \beta} = 0 \qquad & \text{(Stationarity)} \\
h(\beta) = 0 \qquad & \text{(Primal feasibility)} \\
g(\beta) \geq 0 \qquad & \text{(Primal feasibility)} \\
\nu \geq 0 \qquad & \text{(Dual feasibility)} \\
\nu g(\beta) = 0 \qquad & \text{(Complementary slackness).}
\end{aligned}$$

Then solving the KKT system for finding an optimal solution $\hat{\beta}$ can be obtained by using an algorithm such as an interior point method. However, the algorithms to solve the KKT system may not be feasible when a large number of constraints is required as in the SP-GLM. In this thesis we use an alternative method, a Multiplicative Iterative algorithm (Ma, 2010), to handle the inequality constraints. The Multiplicative Iterative algorithm was developed from the KKT necessary conditions to impose the non-negativity constraints, but its computational cost is relatively small.

2.5.3 Multiplicative Iterative algorithm

Our proposed methods in Chapters 3 and 4 apply the Multiplicative Iterative (MI) algorithm (Ma, 2010) to impose the non-negativity constraints. The MI algorithm is a useful iterative algorithm which is able to handle a large number of non-negativity constraints. This MI algorithm only needs the first derivative of the objective function and therefore is easy to implement.

Suppose β is a parameter of interest with constraint $\beta \geq 0$. An objective function is a log-likelihood function $l(\beta)$. We estimate β that is

$$\hat{\beta} = \arg\max_{\beta \geq 0} l(\beta).$$

The corresponding KKT necessary conditions are

$$\frac{\partial l(\beta)}{\partial \beta} = 0 \quad \text{if} \quad \beta > 0,$$
$$\frac{\partial l(\beta)}{\partial \beta} < 0 \quad \text{if} \quad \beta = 0.$$

An optimal solution for β that satisfies non-negativity constraint can be found by solving

$$\beta \frac{\partial l(\beta)}{\partial \beta} = 0, \tag{2.13}$$

subject to $\beta \geq 0$. We rearrange (2.13) to have positive values on both sides

$$-\beta \left[\frac{\partial l(\beta)}{\partial \beta} \right]^{-} = \beta \left[\frac{\partial l(\beta)}{\partial \beta} \right]^{+}, \tag{2.14}$$

where $[a]^{+}$ is a positive component and $[a]^{-}$ is a negative component, so that $[a]^{+} + [a]^{-} = a$. Let $\beta^{(k)}$ be an iterative solution for β at iteration k and define $l'(\beta^{(k)}) = \frac{\partial l(\beta^{(k)})}{\partial \beta}$. This equation (2.14) directs to a temporary update step for the MI algorithm

$$\beta^{\left(k+\frac{1}{2}\right)} = \beta^{(k)} \frac{\left[l'(\beta^{(k)}) \right]^{+}}{-\left[l'(\beta^{(k)}) \right]^{-}}. \tag{2.15}$$

If $\beta^{(k)}$ is positive, the updated value of $\beta^{\left(k+\frac{1}{2}\right)}$ is guaranteed to be positive and $\beta^{\left(k+\frac{1}{2}\right)}$ will only be equal to zero if the numerator of the right hand side of (2.15) is zero. Equation (2.15) can also be seen as a gradient algorithm

$$\beta^{\left(k+\frac{1}{2}\right)} = \beta^{(k)} + \left(\frac{\beta^{(k)}}{-\left[l'(\beta^{(k)})\right]^{-}} \right) l'(\beta^{(k)}). \tag{2.16}$$

Even though this temporary update step can ensure a non-negative updated parameter, it is possible that divergence may occur. The objective function may not increase when moving from $\beta^{(k)}$ to $\beta^{\left(k+\frac{1}{2}\right)}$. A line search step is then included to ensure an increasing objective function.

Line search

Define $\omega^{(k)} \in (0, 1]$ as a step length and $d^{(k)}$ as a search direction. Line search will result in the following iterative scheme

$$\beta^{(k+1)} = \beta^{(k)} + \omega^{(k)} d^{(k)}.$$

Comparing to (2.16), the MI search direction is $d^{(k)} = \left(\frac{\beta^{(k)}}{-\left[l'(\beta^{(k)})\right]^{-}} \right) l'(\beta^{(k)})$.

An optimal line search for maximizing $l(\beta)$ is to find $\omega^{(k)}$ such that

$$l\left(\beta^{(k)} + \omega^{(k)} d^{(k)}\right) = \max_{\omega > 0} l\left(\beta^{(k)} + \omega d^{(k)}\right).$$

However, the exact value of $\omega^{(k)}$ is computationally expensive. In practice, an inexact line search is more desirable since it uses less computation time but has sufficient accuracy. It chooses $\omega^{(k)}$ from an acceptable step size rule to ensure sufficient ascent amount of an objective function in a maximization problem. An example of an inexact line search is Armijo's rule. More details for line search techniques can be found in Luenberger and Ye (1984); Sun and Yuan (2006).

The line search method used in the MI and other algorithms in Chapters 3 and 4 is the Armijo's rule that adopts the following scheme. Suppose ω starts from 1, then we can find ω using the Armijo's condition:

$$l\left(\beta^{(k)}+\omega d^{(k)}\right) \geq l\left(\beta^{(k)}\right)+\gamma\omega d^{(k)} l'\left(\beta^{(k)}\right), \tag{2.17}$$

where $\gamma \in (0,1)$ is a small fixed threshold, e.g. $\gamma = 0.01$. If ω satisfies (2.17), then stop; otherwise reset $\omega = \rho\omega$ where $\rho \in (0,1)$, e.g. $\rho = 0.8$, then re-evaluate (2.17) and continue this procedure until condition (2.17) is satisfied, and then assign ω value to $\omega^{(k)}$.

The MI algorithm with line search can guarantee $l(\beta^{(k+1)}) \geq l(\beta^{(k)})$, and the final update is given by

$$\beta^{(k+1)} = \beta^{(k)} + \omega^{(k)}(\beta^{(k+\frac{1}{2})} - \beta^{(k)}). \tag{2.18}$$

Equation (2.18) is also seen as a gradient algorithm:

$$\beta^{(k+1)} = \beta^{(k)} + \omega^{(k)} \left(\frac{\beta^{(k)}}{-\left[l'(\beta^{(k)})\right]^{-}} \right) l'(\beta^{(k)}). \tag{2.19}$$

2.6 The proposed methods

The SP-GLM is an extension of GLMs where the response mean is a parametric component but the response distribution is partially unspecified and contains a nonparametric component f_0. Exponential tilting is applied to the response distribution so that the model is still written in a full probability setting. The link function is specified similar to the way it is done in GLMs. In this thesis, the SP-GLM only considers the case where the response distribution follows an exponentially tilted distribution with $\phi = w_i = 1$ (see page 9 and 15). The usefulness of the SP-GLM over parametric GLM was already explained in Chapter 1.

The methods proposed in this thesis are efficient model fitting algorithms for the SP-GLM. They are able to handle large data sets with less computational costs. Our approach

distinguishes from the existing methods mainly on the computational algorithm and the asymptotic properties which will be introduced in Chapter 3. Our method can make the SP-GLM more easily applicable in practice.

Our proposed method estimates f_0 together with the regression coefficients. The estimation of f_0 is achieved using a piecewise constant approximation. An iterative approach is used to estimate the regression coefficients and f_0 simultaneously by maximizing the log-likelihood function subject to some constraints. The Lagrange multipliers method and the Multiplicative Iterative algorithm are applied to compute the constrained estimates. The asymptotic properties for the constrained maximum likelihood estimate are provided. Simulation studies are used to test the performance of our proposed method, particularly on accuracy of the parameter estimates and the asymptotic variances. Real data examples are also used to investigate model performance.

In addition, we also propose a special SP-GLM model by adopting the canonical link function in the SP-GLM. The estimation method for this special model is simplified. The main advantage of this special model is that it is simple to use. The regression coefficients and f_0 are again estimated simultaneously via the constrained maximum log-likelihood estimate. The asymptotic properties of parameter estimates are also derived.

3

The semiparametric generalized linear model

3.1 Introduction

The response distribution is only partly specified in the SP-GLM compared with the classical GLM, consequently, regression analysis is more flexible with the SP-GLM. Although the

response distribution is again a member of the exponential family of distributions, the observed data will be used to specify this membership. This approach is very useful in applied settings, especially when there is a doubt about the response distribution. Therefore this approach can help to avoid undesirable results caused by model misspecification such as biased inference for parameters (e.g. Gardner et al. (1995), White (1982), Drake (1993)).

The response density f in this SP-GLM can be written in an exponentially tilted form for the reference density f_0, where f_0 is nonparametric. Since this model is specified by a complete density function, further information about the probability mechanism for generating the data can be obtained. Due to the flexibility of f_0, its exponentially tilted distributions can handle data sets generated by distributions beyond the common distributions for GLMs, such as zero-inflated data and overdispersed count data.

We consider f_0 to be an unknown parameter which will be estimated simultaneously with the regression coefficient parameters $\boldsymbol{\beta}$. However, f_0 is nonparametric and thus is infinite dimensional. Direct estimation of f_0 is ill-conditioned. Some approximation techniques are needed to reduce f_0 to a finite dimensional parameter. Imposing constraints on f_0 also adds complexity to computational algorithms.

To make the model suitable for practical use, we have developed a novel effective computational algorithm (named SP-GLM-I) which can handle large sample sizes and converge with less computational time.

This chapter is set out as follows. The next section presents the approximation for f_0. In Section 3.3, we demonstrate our computational algorithm for model fitting. Section 3.4 presents the asymptotic properties of the estimate. Simulation results are presented in Section 3.5. Finally, conclusions are given in Section 3.6.

3.2 Approximation of the reference density f_0

The SP-GLM with unspecified f_0 was explained in Section 2.3. Recall that the probability density function for the SP-GLM (2.7) is

$$f_i(y|\mathbf{X}_i) \equiv f_i(y) = f_0(y) \exp\{y\theta_i - b(\theta_i; f_0)\},$$

where

$$b(\theta_i; f_0) = \log \int_{\mathcal{Y}} \exp\{y\theta_i\} f_0(y) \mathrm{d}y.$$

However, as shown in Section 2.3.2 that $f_i(y)$ is not identifiable since exponentially tilted versions of f_0 can generate the same $f_i(y)$. The tilted f_0 will change the canonical parameter θ_i by shifting θ_i by a constant θ_0^*. One way to resolve this identifiability issue suggested by Rathouz and Gao (2009) is that f_0 is required to have a specified mean, such as $\int_{\mathcal{Y}} y f_0(y) \mathrm{d}y = \mu_0^*$ where μ_0^* is an arbitrary reference mean within the observed response range. As explained in Section 2.3.2, our strategy (used in this chapter) to make the model identifiable is to fix θ_0^* at zero, and it is clear that there exists a mean value μ_0^* corresponding to this $\theta_0^* = 0$. In other words, we implicitly specified a μ_0^* which corresponds to $\theta_0^* = 0$. Setting θ_0^* to zero is equivalent to considering the y_i's as independent and identically distributed. This makes the corresponding μ_0^* equal to the empirical mean of the observations.

Note that here f_0 is an unspecified density and it will be determined from the data. However, f_0 is an infinite dimensional parameter and this creates computational difficulties. To overcome this problem, we assume f_0 can be expressed as a combination of basis functions:

$$f_0(y) = \sum_{u=1}^{m} \alpha_u \psi_u(y). \tag{3.1}$$

Here $\psi_u(y)$ are some known non-negative basis functions and α_u are the coefficients for these basis functions. The α_u are constrained to be non-negative to ensure that $f_0(y)$ is non-negative. In general, we require $m \leq n$ where n is the sample size.

In this thesis, we consider particularly the indicator basis functions for $\psi_u(y)$. Equivalently, the approximate f_0 becomes a piecewise constant function. Suppose the observed response vector $y = [y_1 \dots y_n]^T$ has minimum and maximum observations $y_{(1)}$ and $y_{(n)}$, respectively. The range of the observed response is $\mathcal{H} = \left[y_{(1)}, y_{(n)}\right]$. Indicator basis functions partition $\mathcal{H}$ into m "bins" which are denoted by $B_1, \dots, B_m$. The methods of equal-width, equal-frequency, or fixed-frequency can be applied for selecting the bin sizes. The bins are designed to contain a certain number of observations, so that the location of bins will change from one sample to another. The partitions of the bins are mutually exclusive and exhaustive. So if $y \in B_u$, we have $\psi_u(y) = 1$ and otherwise $\psi_u(y) = 0$. Then $f_0(y)$ in (3.1) becomes $f_0(y) = \alpha_u$ for $y \in B_u$. Let $p_u = \alpha_u \delta_u$ where δ_u is the width of bin B_u. Since f_0 is a density, it must satisfy the constraints: (i) $f_0(y) \geq 0$ and (ii) $\int_{\mathcal{Y}} f_0(y)\mathrm{d}y = 1$. So under indicator basis functions, p_u are required to be: (i) $p_u \geq 0$ for $u = 1, \dots, m$ and (ii) $\sum_{u=1}^{m} p_u = 1$. Thus p_u themself can be seen as probability mass.

3.2.1 Log-likelihood function

For the probability density function of the SP-GLM in (2.7) and the approximation of f_0, the log-likelihood function becomes

$$l(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} (y_i \theta_i - b(\theta_i; \boldsymbol{p})) + \sum_{u=1}^{m} n_u \log p_u, \tag{3.2}$$

where n_u is the number of observed responses in bin B_u. A direct derivation yields

$$b(\theta_i; \boldsymbol{p}) = \log \sum_{u=1}^{m} p_u \int_{B_u} \exp\{y\theta_i\}\, \mathrm{d}y = \log \sum_{u=1}^{m} \frac{p_u}{\theta_i} \left[\exp\{\max(B_u)\theta_i\} - \exp\{\min(B_u)\theta_i\}\right]. \tag{3.3}$$

Note that if c_u is the mid-point of bin B_u and is used to be the representative value of each bin, then $b(\theta_i; \boldsymbol{p})$ can be approximated by:

$$b(\theta_i; \boldsymbol{p}) = \log \sum_{u=1}^{m} \exp\{c_u \theta_i\} p_u. \tag{3.4}$$

This formula (3.4) is a simple version of formula (3.3) and is used for calculation.

From an estimation point of view, there is no difference between p_u or α_u. Along with the log-likelihood function (3.2), constraints (i) - (ii) on p_u defined above must be imposed. Technically, only $(m-2)$ p_u's are free. For identifiability, the mean constraint on p_u that is $\sum_{u=1}^{m} c_u p_u = \mu_0^*$, is applied by implicitly specifying μ_0^* corresponding to $\theta_0^* = 0$ (see Sections 2.3.2 and 3.2).

3.3 Constrained maximum likelihood estimation

We apply a special constrained maximum likelihood estimation method to estimate $\boldsymbol{\beta}$ and $\boldsymbol{p}$ simultaneously. The non-negativity of p_u are constrained directly using the MI (Ma, 2010) algorithm and the method of Lagrange multipliers is applied to impose the equality constraint. Suppose λ is the Lagrange multiplier. The Lagrangian for the equality constraint is

$$\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} (y_i \theta_i - b(\theta_i; \boldsymbol{p})) + \sum_{u=1}^{m} n_u \log p_u - \lambda \left(1 - \sum_{u=1}^{m} p_u\right). \tag{3.5}$$

We wish to maximize $\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{p})$ subject to $p_u \geq 0$. The corresponding KKT conditions for this constrained optimization are,

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\beta}} = 0, \quad \frac{\partial \mathscr{L}}{\partial p_u} = 0 \text{ if } p_u > 0, \quad \frac{\partial \mathscr{L}}{\partial p_u} < 0 \text{ if } p_u = 0, \quad \text{and} \quad \frac{\partial \mathscr{L}}{\partial \lambda} = 0.$$

Note that the last equation gives the original equality constraint.

We follow the Newton - MI algorithm of Ma, Heritier, and Lô (2014) to develop an iterative method, named the MI - Scoring algorithm, to estimate $\boldsymbol{p}$ and $\boldsymbol{\beta}$. Suppose the estimate value

of any parameter a at iteration k is denoted by $a^{(k)}$. We will develop an alternating algorithm where $\boldsymbol{p}$ and $\boldsymbol{\beta}$ are updated separately in each iteration. We adopt the strategy that at each iteration, we first apply the MI algorithm to update $\boldsymbol{p}^{(k+1)}$ using $(\boldsymbol{\beta}^{(k)}, \boldsymbol{p}^{(k)}, \boldsymbol{\theta}^{(k)})$. Note that θ_i's are functions of $\boldsymbol{\beta}$ and $\boldsymbol{p}$, all θ_i and $b(\theta_i, \boldsymbol{p})$ need to be updated first after obtaining the updated values of $\boldsymbol{p}$, and then they will be updated again after updating $\boldsymbol{\beta}$. Let $\boldsymbol{\theta}^{(k+\frac{1}{2})}$ be the estimate of $\boldsymbol{\theta}$ based on $\boldsymbol{p}^{(k+1)}$ and $\boldsymbol{\beta}^{(k)}$. After that we use $(\boldsymbol{\beta}^{(k)}, \boldsymbol{p}^{(k+1)}, \boldsymbol{\theta}^{(k+\frac{1}{2})})$ to update $\boldsymbol{\beta}^{(k+1)}$ by applying the Fisher scoring algorithm, then $\boldsymbol{\theta}^{(k+1)}$ is obtained from $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{p}^{(k+1)}$. This process is repeated until convergence occurs with the convergence criterion defined to be $\max_u |p_u^{(k+1)} - p_u^{(k)}| < 10^{-6}$ and $\max_j |\beta_j^{(k+1)} - \beta_j^{(k)}| < 10^{-6}$. Details of the algorithm are given as follows.

3.3.1 Estimate the p_u's

The MI algorithm is applied to estimate $\boldsymbol{p} \geq 0$. This is achieved by solving the KKT conditions:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial p_u} = 0 & \text{if} \quad p_u > 0, \\ \dfrac{\partial \mathscr{L}}{\partial p_u} < 0 & \text{if} \quad p_u = 0. \end{cases}$$

A necessary condition for these equations is:

$$p_u \frac{\partial \mathscr{L}}{\partial p_u} = 0 \quad \text{subject to} \quad p_u \geq 0.$$

For our problem, the score function for p_u is

$$\frac{\partial \mathscr{L}}{\partial p_u} = -\sum_{i=1}^{n} (y_i - b'(\theta_i; \boldsymbol{p}))(c_u - b'(\theta_i; \boldsymbol{p})) \frac{1}{b''(\theta_i; \boldsymbol{p})} \exp\{c_u\theta_i - b(\theta_i; \boldsymbol{p})\} + \frac{n_u}{p_u} + \lambda, \quad (3.6)$$

where

$$b'(\theta_i; \boldsymbol{p}) = \sum_{u=1}^{m} c_u p_u \exp\{c_u\theta_i - b(\theta_i; \boldsymbol{p})\}, \quad (3.7)$$

$$b''(\theta_i; \boldsymbol{p}) = \sum_{u=1}^{m} (c_u - b'(\theta_i; \boldsymbol{p}))^2 p_u \exp\{c_u\theta_i - b(\theta_i; \boldsymbol{p})\}. \quad (3.8)$$

From $\sum_{u=1}^{m} p_u \frac{\partial \mathscr{L}}{\partial p_u} = 0$, we can easily derive $\lambda = -n$.

We separate the score function (3.6) into positive and negative terms, and then rearrange the equation so that its both sides are positive. This gives us a temporary updating formula:

$$p_u^{(k+\frac{1}{2})} = p_u^{(k)} \left(\frac{\frac{n_u}{p_u^{(k)}} - \sum_{i=1}^{n} [A]^{-} \left(\frac{1}{b''(\theta_i^{(k)}; \boldsymbol{p}^{(k)})} \right) \exp\left\{ c_u \theta_i^{(k)} - b(\theta_i^{(k)}; \boldsymbol{p}^{(k)}) \right\} + \varepsilon}{n + \sum_{i=1}^{n} [A]^{+} \left(\frac{1}{b''(\theta_i^{(k)}; \boldsymbol{p}^{(k)})} \right) \exp\left\{ c_u \theta_i^{(k)} - b(\theta_i^{(k)}; \boldsymbol{p}^{(k)}) \right\} + \varepsilon} \right) \tag{3.9}$$

where $A = \left(y_i - b'(\theta_i^{(k)}; \boldsymbol{p}^{(k)}) \right) \left(c_u - b'(\theta_i^{(k)}; \boldsymbol{p}^{(k)}) \right)$. In this scheme, ε is a small positive constant added to avoid division by zero, and it has no effect on the estimated value of p_u. $p_u^{(k+\frac{1}{2})}$ in (3.9) is clearly non-negative if $p_u^{(k)} > 0$. However, $\mathscr{L}\left(\boldsymbol{\beta}^{(k)}, \boldsymbol{p}\right)$ may not be guaranteed to increase when it moves from $\boldsymbol{p}^{(k)}$ to $\boldsymbol{p}^{(k+\frac{1}{2})}$. Thus in the second step we require a line search step size $\omega_1^{(k)} \in (0, 1]$ to guarantee that $\mathscr{L}\left(\boldsymbol{\beta}^{(k)}, \boldsymbol{p}^{(k+1)}\right) \geq \mathscr{L}\left(\boldsymbol{\beta}^{(k)}, \boldsymbol{p}^{(k)}\right)$. Therefore p_u is updated by

$$p_u^{(k+1)} = p_u^{(k)} + \omega_1^{(k)} \left(p_u^{(k+\frac{1}{2})} - p_u^{(k)} \right). \tag{3.10}$$

After obtaining $\boldsymbol{p}^{(k+1)}$ we need to update $\boldsymbol{\theta}$ to give $\boldsymbol{\theta}^{(k+\frac{1}{2})}$. The estimation algorithm for $\boldsymbol{\theta}$ is explained in the next section.

3.3.2 Computation of $\boldsymbol{\theta}$

As explained in Chapter 2, in parametric GLMs, parameters θ_i are explicitly known from the distribution. In the SP-GLM, however, θ_i are unknown but can be derived from the mean condition:

$$\mu_i = b'(\theta_i; \boldsymbol{p}). \tag{3.11}$$

There is a restriction on the possible range of the mean function $b'(\theta_i; \boldsymbol{p})$. Recall that c_u is a value for bin B_u, then possible upper and lower bounds for $b'(\theta_i; \boldsymbol{p})$ (see equation (3.7)) is $[c_{(1)}, c_{(m)}]$ where $c_{(1)} = \min\{c_1, \ldots, c_m\}$ and $c_{(m)} = \max\{c_1, \ldots, c_m\}$. Thus when we estimate

θ_i, if the corresponding μ_i value is outside these bounds, we need to reset μ_i to the boundary value $c_{(1)}$ or $c_{(m)}$.

If $\boldsymbol{\beta}$ and $\boldsymbol{p}$ are known, we can obtain θ_i by solving equation (3.11). The Newton algorithm can be used to estimate θ_i. Note that $\mu_i^{(k)} = g^{-1}(\mathbf{X}_i\boldsymbol{\beta}^{(k)})$. For fixed $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{p}^{(k+1)}$, we update θ_i according to equation (3.12)

$$\theta_i^{(k,r+1)} = \theta_i^{(k,r)} - \omega_{2i}^{(r)} \left[b''(\theta_i^{(k,r)}; \boldsymbol{p}^{(k+1)}) \right]^{-1} \left(b'(\theta_i^{(k,r)}; \boldsymbol{p}^{(k+1)}) - \mu_i^{(k)} \right), \tag{3.12}$$

where $\omega_{2i}^{(r)} \in [0, 1]$ is a step length to assure convergence of this algorithm. There is a computational issue when the absolute value of $\exp\{c_u\theta_i\}$ becomes too large and may exceed the largest positive floating-point number that MATLAB can handle. Based on our simulation results, the best strategy is to control the increment on θ_i at each iteration. For each iteration, we have $\omega_{2i}^{(r)} = 10^{-d_i}$ where d_i is an integer digits of the increment on θ_i. This strategy can work well to find θ_i that makes μ_i and $b'(\theta_i; \boldsymbol{p})$ equivalent. The last term $(b'(\theta_i; \boldsymbol{p}) - \mu_i)$ is monotonic increasing with respect to θ_i. The convergence criterion of this algorithm is set as $\max_i |\theta_i^{(k,r+1)} - \theta_i^{(k,r)}| < 10^{-6}$ and $\max_i |b'(\theta_i^{(k,r+1)}; \boldsymbol{p}^{(k+1)}) - \mu_i^{(k)}| < 10^{-6}$. Note that this equation for θ_i must be "fully solved" i.e. at each iteration, the updated θ_i must be iterated until it converges, to give $\boldsymbol{\theta}^{(k+\frac{1}{2})}$, as otherwise computations for other components can become unstable. After updating $\boldsymbol{\theta}$ we next need to update the estimate for $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}^{(k+1)}$. Then $\boldsymbol{\theta}$ is updated again using the same formula as (3.12) but now based on $(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{p}^{(k+1)})$.

3.3.3 Estimating the regression coefficient $\boldsymbol{\beta}$

In the SP-GLM, the linear predictor η_i, the link function $g(\cdot)$ and the response mean μ_i are defined in the same way as in the GLMs (see Sections 2.1.2 and 2.1.3). The linear predictor is defined as $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$. The linear predictor and the response mean are linked through the specified link function that is $g(\mu_i) = \eta_i$. Define $h(\cdot) = g^{-1}(\cdot)$ as the inverse link function,

and thus $\mu_i = h(\eta_i)$. The coefficient parameters $\boldsymbol{\beta}$ can be estimated using the Fisher scoring algorithm. The score function for $\boldsymbol{\beta}$ is

$$S_{\boldsymbol{\beta}} = \frac{\partial \mathscr{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{b''(\theta_i; \boldsymbol{p})} h'(\eta_i) \mathbf{X}_i,$$

where $h'(\eta_i) = \partial \mu_i / \partial \eta_i$. The Fisher information matrix for $\boldsymbol{\beta}$ is given by

$$- \mathrm{E}\left(\frac{\partial^2 \mathscr{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right) = \sum_{i=1}^{n} \mathbf{X}_i^T \frac{1}{b''(\theta_i; \boldsymbol{p})} [h'(\eta_i)]^2 \mathbf{X}_i. \tag{3.13}$$

Define $\mathbf{y}$ to be n-column vector for all y_i. Using the Fisher scoring algorithm, $\boldsymbol{\beta}$ is updated by:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \omega_3^{(k)} \left(\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{V}^{(k)} \left(\mathbf{y} - \boldsymbol{\mu}^{(k)}\right), \tag{3.14}$$

where $\mathbf{W}^{(k)} = \mathrm{diag}\left(h'(\eta_i^{(k)})[b''(\theta_i^{(k+\frac{1}{2})}; \boldsymbol{p}^{(k+1)})]^{-1} h'(\eta_i^{(k)})\right)$ and $\mathbf{V}^{(k)} = \mathrm{diag}\Big([b''(\theta_i^{(k+\frac{1}{2})}; \boldsymbol{p}^{(k+1)})]^{-1} h'(\eta_i^{(k)})\Big)$. Here $\omega_3^{(k)} \in (0, 1]$ is a line search step size applied to ensure the increasing log-likelihood $\mathscr{L}\left(\boldsymbol{\beta}, \boldsymbol{p}^{(k+1)}\right)$ when $\boldsymbol{\beta}$ moves from $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+1)}$.

Once this updated $\boldsymbol{\beta}^{(k+1)}$ is obtained, in order to get the estimated log-likelihood, we need to obtain $\boldsymbol{\theta}^{(k+1)}$ by solving the Newton algorithm (3.12) again using $(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{p}^{(k+1)})$. If this updated $\boldsymbol{\beta}^{(k+1)}$ does not increase the log-likelihood, a line search step size $\omega_3^{(k)}$ is applied to reupdate $\boldsymbol{\beta}^{(k+1)}$ and then reupdate $\boldsymbol{\theta}^{(k+1)}$ until the log-likelihood increases, i.e. $\mathscr{L}\left(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{p}^{(k+1)}\right) \geq \mathscr{L}\left(\boldsymbol{\beta}^{(k)}, \boldsymbol{p}^{(k+1)}\right)$. Note that equality holds here only when the iterations have converged.

3.4 Asymptotic properties

The asymptotic properties of the constrained maximum likelihood estimation of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{p}}$ are given in this section for a fixed number of basis functions i.e. $\boldsymbol{p}$ is finite-dimensional. Let the combined parameter vector be $\boldsymbol{\gamma} = \left(\boldsymbol{\beta}^T, \boldsymbol{p}^T\right)^T$, which has dimension $q + m$. The

log-likelihood function (3.2) is then written as $l(\boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{p})$. Let $\boldsymbol{\gamma}^*$ be the true parameter value of $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$ be the maximum likelihood estimate of $\boldsymbol{\gamma}$. $\hat{\boldsymbol{\gamma}}$ is estimated by maximizing $l(\boldsymbol{\gamma})$ subject to constraint $p_u \geq 0$ and $\sum_{u=1}^{m} p_u = 1$.

Suppose the Fisher information matrix at $\boldsymbol{\gamma}$ is given by

$$I(\boldsymbol{\gamma}) = \begin{bmatrix} I_{\boldsymbol{\beta}\boldsymbol{\beta}} & I_{\boldsymbol{\beta}\boldsymbol{p}} \\ I_{\boldsymbol{p}\boldsymbol{\beta}} & I_{\boldsymbol{p}\boldsymbol{p}} \end{bmatrix}$$

where components of this matrix are defined as follows:

$I_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is the Fisher information matrix for $\boldsymbol{\beta}$ with $q \times q$ dimensions similar to that is defined in (3.13). $I_{\boldsymbol{\beta}\boldsymbol{p}} = -\mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{p}}\right)$ is the $q \times m$ Fisher information matrix for the covariance of $\boldsymbol{\beta}$ and $\boldsymbol{p}$ with the elements corresponding to p_u given by

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial p_u}\right) = -\sum_{i=1}^{n} (c_u - b'(\theta_i; \boldsymbol{p})) \exp\{c_u\theta_i - b(\theta_i; \boldsymbol{p})\} p_u \frac{1}{b''(\theta_i; \boldsymbol{p})} h'(\eta_i)\mathbf{X}_i.$$

$I_{\boldsymbol{p}\boldsymbol{\beta}}$ is the transpose of a matrix $I_{\boldsymbol{\beta}\boldsymbol{p}}$ with $m \times q$ dimensions. $I_{\boldsymbol{p}\boldsymbol{p}}$ is the $m \times m$ Fisher information matrix for $\boldsymbol{p}$ with the elements given by

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial p_u \partial p_v}\right) = \sum_{i=1}^{n} (c_u - b'(\theta_i; \boldsymbol{p}))(c_v - b'(\theta_i; \boldsymbol{p})) \exp\{(c_u + c_v)\theta_i - 2b(\theta_i; \boldsymbol{p})\} p_u p_v \frac{1}{b''(\theta_i; \boldsymbol{p})} + \frac{n_u}{p_u^2} \mathbf{1}_{u=v},$$

where $\mathbf{1}_{u=v}$ is an indicator for $u = v$ (i.e. diagonal element of $I_{\boldsymbol{p}\boldsymbol{p}}$).

For constrained maximum likelihood problems, the standard asymptotic variance from the inverse of the Fisher information matrix (for unconstrained problem) will provide incorrect variances since the active constraints will affect the variance.

For the SP-GLM, there are active constraints on $\boldsymbol{p}$. We need to consider the possibility of active constraints on $\boldsymbol{p}$, especially when we want to do inference about $\boldsymbol{p}$. For example, an insurance company may be interested in the probability that a particular person will claim at least once, i.e. that the number of claims ≥ 1, for the next year, then $\boldsymbol{p}$ is required when computing this probability.

The asymptotic properties developed here take account of the active constraints on $\boldsymbol{p}$. We closely follow the asymptotic properties presented in Moore, Sadler, and Kozick (2008). The asymptotic properties for semiparametric proportional hazard models developed by Ma, Couturier, Heritier, and Marschner (2017) are also considered as a guide for the asymptotic properties in this semiparametric model.

Next we explain the active constraints. Our model deals with the equality constraint $\sum_{u=1}^{m} p_u - 1 = 0$ and the inequality constraints $p_u \geq 0$ for $u = 1, \ldots, m$. The equality constraint is always active. While an inequality constraint is active if $p_u = 0$ and $\frac{\partial \mathscr{L}}{\partial p_u} < 0$. Suppose there are m_1 active inequality constraints. Then there are $m_1 + 1$ active constraints in total.

Let the first derivative of the active constraints with respect to $\boldsymbol{\gamma}$ be denoted as matrix $G(\boldsymbol{\gamma}) = [\mathbf{0}^T_{(m_1+1)\times q}, G_1(\boldsymbol{p})^T_{(m_1+1)\times m}]^T$ with dimension $(m_1 + 1) \times (q + m)$ where $\mathbf{0}$ is a zero matrix. The zero matrix in $G(\boldsymbol{\gamma})$ comes from the derivative of the active constraints with respect to $\boldsymbol{\beta}$ and matrix $G_1(\boldsymbol{p}) = [\mathbf{1}^T_{1\times m}, G_2(\boldsymbol{p})^T_{m_1\times m}]^T$ is the derivative result of the active constraints with respect to $\boldsymbol{p}$. The row of 1s corresponds to the derivative of the equality constraint. Matrix $G_2(\boldsymbol{p})$ is a matrix of 0s and 1s where each row only has the value of 1 at the column u if $p_u = 0$ and $\frac{\partial \mathscr{L}}{\partial p_u} < 0$. To be clear, let $\tau = [\tau_1, \ldots, \tau_m]^T$ be an indicator vector of the active inequality constraints that is $\tau_u = 1$ for the corresponding active constraint and zero otherwise. Then the rows of matrix $G_2(\boldsymbol{p})$ are selected from an identity matrix $\mathbf{I}_{m\times m}$ using the non-zero portion of τ.

We define matrix $U(\boldsymbol{\gamma})_{(q+m)\times(q+m-m_1-1)}$ whose columns form an orthonormal basis of the null space of $G(\boldsymbol{\gamma})$ such that

$$G(\boldsymbol{\gamma})U(\boldsymbol{\gamma}) = \mathbf{0}_{(m_1+1)\times(q+m-m_1-1)} \quad \text{and} \quad U^T(\boldsymbol{\gamma})U(\boldsymbol{\gamma}) = \mathbf{I}_{(q+m-m_1-1)\times(q+m-m_1-1)}. \tag{3.15}$$

Then the inverse of the Fisher information matrix for $\boldsymbol{\gamma}$ accommodating active constraints is

$$F(\boldsymbol{\gamma})^{-1} = U(\boldsymbol{\gamma})(U^T(\boldsymbol{\gamma})I(\boldsymbol{\gamma})U(\boldsymbol{\gamma}))^{-1}U^T(\boldsymbol{\gamma}). \tag{3.16}$$

The assumptions required for consistency in Theorem 1 and the asymptotic normality in Theorem 2 are:

A1. Random variables $\mathbf{X}_i$ for $i = 1, \ldots, n$ are independent and identically distributed.

A2. The distribution of $\mathbf{X}_i$ is independent of $\boldsymbol{\gamma}$.

A3. The domain of $\boldsymbol{\gamma}$, denoted by Γ, is a compact subset of R^{q+m}.

A4. $\mathrm{E}_{\gamma^*}[n^{-1}l(\boldsymbol{\gamma})]$ exists and has a unique maximum at $\boldsymbol{\gamma}^* \in \Gamma$.

A5. $l(\boldsymbol{\gamma})$ is continuous over Γ and is twice differentiable in a neighborhood of $\boldsymbol{\gamma}^*$.

A6. The Fisher information matrix at $\boldsymbol{\gamma}^*$ exists.

Note that by keeping m fixed, the true data generating process is assumed to have a density created by exponential tilts from a piecewise constant density, for which there is a true $\boldsymbol{p}$ towards which the estimate converges.

Theorem 1 *Assume that assumptions A1 - A6 hold, the constrained maximum likelihood estimation of $\hat{\boldsymbol{\gamma}}$ is consistent for $\boldsymbol{\gamma}^*$ as $n \to \infty$.*

Theorem 2 *Assume that assumptions A1 - A6 hold and assume that there are m_1 active inequality constraints in the maximum likelihood estimate of $\boldsymbol{p}$. Let matrix $U(\boldsymbol{\gamma}^*)$ be defined as in (3.15). Then the distribution of $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$ converges in distribution to the multivariate normal distribution $N(0_{(q+m)\times 1}, F(\boldsymbol{\gamma}^*)^{-1})$ as $n \to \infty$.*

Proof. The proof is omitted here as this is a simple modification of that given in Moore et al. (2008) and Ma et al. (2017).

Our asymptotic standard error results are compared with standard errors obtained from Monte Carlo simulations in Section 3.5. These results demonstrate that our asymptotic variances on $\boldsymbol{\beta}$ are accurate.

3.5 Simulation results

Simulation studies were conducted to investigate the effectiveness of our proposed iterative method (SP-GLM-I) to fit the SP-GLM with response data that are generated from distributions from both within and beyond the standard distributions of GLM framework. We examine the performance of SP-GLM-I in handling datasets with small, medium and large sample sizes. The regression coefficient estimates and inferences are also examined. We also investigate the accuracy of the asymptotic variance proposed in Section 3.4. We perform three simulation studies and set the goals for these simulation studies as follows. We expect that when the response distribution is generated by a traditional GLM, the SP-GLM-I algorithm will produce results close to the standard GLM. When the response distribution is not generated by the standard GLM, SP-GLM-I performs much better than the standard GLM. In addition, we expect that for small and medium size data sets, SP-GLM-I is as good as the existing SP-GLM method of Huang (2014) (which we will hereafter call SP-GLM-H). We also expect that for large data sets, SP-GLM-H will not produce results since it cannot handle large data sets, while the SP-GLM-I method is still feasible. We compare these simulation results using biases and variances, and also using Type I error rates.

In the first simulation, we aim to investigate the performance of SP-GLM-I to fit the response data that are generated from a standard distribution in the GLM framework, that is the Poisson distribution. We use small sample sizes ($n = 30$) and we wish to explore the effectiveness of our model fitting method compared with the standard GLM and SP-GLM of Huang.

In the second simulation, the performance of SP-GLM-I is examined where the response data are generated from a zero-inflated Poisson distribution, a non-standard distribution in the GLM framework. In this simulation, sample sizes of $n = 300$ and $n = 10,000$ are used to

explore the performance of SP-GLM-I.

In the third simulation, we investigate the performance of SP-GLM-I with different numbers of bins m. The response data in this simulation are generated from the exponential distribution with inverse link. Note that we only investigate the choice of m for continuous responses. When the response is count data as in the first and second simulations, we choose m equal to the number of possible values for that response variable.

For each simulation, the results from SP-GLM-I are compared to the results from the SP-GLM-H and the classical GLM methods. The MATLAB code for our SP-GLM-I method is provided in Appendix A.1. We compare SP-GLM-I with the SP-GLM-H method to evaluate the accuracy of SP-GLM-I for data sets with small and medium sizes. We use the `spglm4()` MATLAB function provided by Huang to implement SP-GLM-H. This function uses the built-in MATLAB solver function `fmincon()` to solve the optimization problem. For GLM, we use the `fitglm()` MATLAB function. We compare the results of SP-GLM-I with the classical GLM to investigate the accuracy of SP-GLM-I when fitting standard GLM models.

To assess the accuracy of SP-GLM-I by using simulation studies, we compare its average estimate (MEAN), bias (BIAS), average asymptotic standard error (AASE), Monte Carlo standard error (MCSE), and mean squared error (MSE) on $\hat{\boldsymbol{\beta}}$ with other methods. More specifically, for each simulation study we generate N samples of size n. For each sample of size n, we fit the model and obtain the estimates of $\boldsymbol{\beta}$ and their corresponding asymptotic standard errors for each method. Define $\hat{\beta}_{jr}$ for $j = 0, 1, 2$ and $r = 1, \ldots, N$ to be the estimate of β_j for the rth sample and let $\text{ASE}\left(\hat{\beta}_{jr}\right)$ be the asymptotic standard error of $\hat{\beta}_{jr}$. Then we can compute

$$\text{MEAN}\left(\hat{\beta}_j\right) = \frac{\sum_{r=1}^{N} \hat{\beta}_{jr}}{N}, \tag{3.17}$$

$$\text{BIAS}\left(\hat{\beta}_j\right) = \beta_j - \text{MEAN}\left(\hat{\beta}_j\right), \tag{3.18}$$

$$\text{AASE}\left(\hat{\beta}_j\right) = \frac{\sum_{r=1}^{N} \text{ASE}\left(\hat{\beta}_{jr}\right)}{N}, \tag{3.19}$$

$$\text{MCSE}\left(\hat{\beta}_j\right) = \sqrt{\frac{1}{N-1}\sum_{r=1}^{N}\left[\hat{\beta}_{jr} - \text{MEAN}\left(\hat{\beta}_j\right)\right]^2}, \tag{3.20}$$

and

$$\text{MSE}\left(\hat{\beta}_j\right) = \left[\text{BIAS}\left(\hat{\beta}_j\right)\right]^2 + \left[\text{AASE}\left(\hat{\beta}_j\right)\right]^2. \tag{3.21}$$

The average asymptotic standard errors are compared with the Monte Carlo standard errors to examine the accuracy of our asymptotic variances. To avoid confusion, the MCSE results are shown in brackets in Tables 3.1, 3.4, 3.7 and 3.8. The asymptotic covariance matrix $F(\boldsymbol{\gamma}^*)^{-1}$ for SP-GLM-I is computed as specified in equation (3.16). The SP-GLM-I asymptotic standard error values for $\hat{\boldsymbol{\beta}}$ are the square root of the first q diagonal values in $F(\boldsymbol{\gamma}^*)^{-1}$. The asymptotic standard errors of $\hat{\boldsymbol{\beta}}$ for SP-GLM-H are not available from Huang's `spglm4()` MATLAB function, so we have modified the `spglm4()` function to include asymptotic standard errors for $\hat{\boldsymbol{\beta}}$ from the square root of the inverse of the Fisher information matrix for $\boldsymbol{\beta}$ (Huang, 2014, Proposition 1).

To explore the performance of SP-GLM-I, the average residual sum of squares of different methods are also compared. Let y_{ir} be the ith response value for the rth sample and $\hat{\mu}_{ir}$ be the corresponding fitted value. The residual sum of squares for the rth sample (RSS_r) is

$$\text{RSS}_r = \sum_{i=1}^{n} \left[y_{ir} - \hat{\mu}_{ir}\right]^2. \tag{3.22}$$

Then the average residual sum of squares (ARSS) is given by

$$\text{ARSS} = \frac{\sum_{r=1}^{N} \text{RSS}_r}{N}. \tag{3.23}$$

To examine the computation costs of each computational method, we compute the average computational time (in seconds) for each method. Define TIME_r to be the computational

time spend for the rth sample. Then the average computational time spend (ATIME) is given by

$$\text{ATIME} = \frac{\sum_{r=1}^{N} \text{TIME}_r}{N}. \tag{3.24}$$

To investigate the accuracy of the parameter inference of the SP-GLM-I algorithm, we perform a hypothesis test. We calculate Type I errors for testing the null hypothesis at 5% significance level using SP-GLM-I and compare them with those of other methods. The response data are generated under the null hypothesis and Type I error rates are calculated from:

(i) the Wald test with average asymptotic standard error of $\hat{\beta}_1$ (WALD-AASE),

(ii) the Wald test with Monte Carlo standard error of $\hat{\beta}_1$ (WALD-MCSE) and

(iii) the likelihood ratio test for chi-square distribution with 1 degrees of freedom (LR).

3.5.1 Log-linear model

In this section, simulated count data with log-linear model is used to examine the performance of the model fitting for the SP-GLM-I method. In this simulation, values for x_{i1} and x_{i2} were generated independently from the uniform (0,1) distribution. We simulated $N = 1,000$ samples of size $n = 30$. The true $\boldsymbol{\beta}$ was $[1, -0.3, 0.5]^T$. In each sample, the response variable was randomly generated from a Poisson distribution with the log link. That is

$$y_i \sim \text{Poi}\left(\exp\{1 - 0.3x_{i1} + 0.5x_{i2}\}\right).$$

In the model fitting, the Poisson distribution with log link was specified for the GLM model fitting; but SP-GLM-I and SP-GLM-H only require us to specify a link function which was the log link in these cases.

The simulation results in Table 3.1 suggest that SP-GLM-I provides accurate regression coefficient estimation since the coefficient biases and the MSEs are small and close to those of GLM. Furthermore, the AASE values of SP-GLM-I are close to the MCSE values and they are also close to the standard errors of SP-GLM-H and GLM. This indicates the accuracy of the asymptotic variances presented in Section 3.4.

Table 3.1: Simulation results for Poisson response with $n = 30$.

Method		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
True $\boldsymbol{\beta}$		1.0000	-0.3000	0.5000
SP-GLM-I	MEAN	0.9990	-0.3137	0.4810
	BIAS	-0.0010	-0.0137	-0.0190
	AASE	0.3044	0.3299	0.3874
	MCSE	(0.3080)	(0.3574)	(0.4012)
	MSE	0.0926	0.1090	0.1505
SP-GLM-H	MEAN	0.9981	-0.3149	0.4832
	BIAS	-0.0019	-0.0149	-0.0168
	AASE	0.2859	0.3103	0.3655
	MCSE	(0.3085)	(0.3594)	(0.4033)
	MSE	0.0817	0.0965	0.1339
GLM	MEAN	1.0007	-0.3126	0.4759
	BIAS	0.0007	-0.0126	-0.0241
	AASE	0.3093	0.3404	0.3973
	MCSE	(0.3054)	(0.3536)	(0.3979)
	MSE	0.0957	0.1161	0.1584

The ARSS and ATIME are presented in Table 3.2. The ARSS of SP-GLM-I is similar to the results of the SP-GLM-H and GLM methods, indicating SP-GLM-I provides good fit for small data sets ($n = 30$). In addition, SP-GLM-I used less computational time on model

Table 3.2: ARSS and ATIME (in seconds) for the model fitting of Poisson response with $n = 30$.

Method	ARSS	ATIME
SP-GLM-I	76.44	0.028
SP-GLM-H	76.45	0.068
GLM	76.40	0.002

fitting than SP-GLM-H.

Type I error

Next we used the previous data set for hypothesis testing, H_0: $\beta_1 = -0.3$ and H_1: $\beta_1 \neq -0.3$ and calculated Type I error rates. Type I errors were tested based on the Wald test with average asymptotic standard error and with Monte Carlo standard error of $\hat{\beta}_1$, and also the likelihood ratio test at 5% significance level. Results are shown in Table 3.3. Type I error rates of SP-GLM-I from the WALDs and LR are all accurate as they are relatively close to 0.05, and also the results are close to those of GLM. This suggests that the parameter inference for the SP-GLM-I method is accurate. Note that Type I error rates of WALD-AASE and LR of SP-GLM-H are higher than the nominated value 0.05.

Table 3.3: Type I errors for β_1 for Poisson response with $n = 30$.

Method	WALD-AASE	WALD-MCSE	LR
SP-GLM-I	0.069	0.054	0.055
SP-GLM-H	0.103	0.054	0.086
GLM	0.062	0.048	0.064

In conclusion, the SP-GLM-I method can fit the log-linear Poisson model well and it can provide accurate regression coefficient estimates and inferences.

3.5.2 Zero-inflated data

Zero-inflated data is data that contains an excessive frequency of zeros, and thus the probability mass at zero exceeds that allowed by, for example, the Poisson distribution. Zero-inflated data is regularly found in insurance claim applications. The number of claims for an insurance policy can follow a zero-inflated Poisson distribution or a zero-inflated negative binomial distribution. For example, the majority of the policies for car insurance make no claim. Some of them may make a few claims during a year. It is rare for policies to generate more than 4 claims a year.

Simulation studies of zero-inflated Poisson data were conducted to assess the performance of SP-GLM-I. The response data were generated from a mixture distribution. The zero-inflated Poisson distribution contains two sub-populations:

(i) sub-population 1 is the zero counts with proportion π,

(ii) sub-population 2 is the counts that follow the Poisson distribution with proportion $(1-\pi)$.

For data generation, first we generated a Bernoulli random variable to define a sub-population. Then the response count Y_i was generated corresponding to this sub-population. Thus Y_i had its probability function given by

$$f(y_i) = \begin{cases} \pi + (1-\pi)e^{-\mu_i} & \text{for} \quad y_i = 0 \\ (1-\pi)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & \text{for} \quad y_i = 1, 2, \ldots . \end{cases}$$

For this simulation study, we specified $\boldsymbol{\beta} = [1, -0.3, 0.5]^T$. The zero counts proportion $\pi = 0.3$ was applied. Note that the constant zero-inflated Poisson model is not an exponential family, thus such a data-generating mechanism is in fact misspecified for the SP-GLM. The sample sizes of $n = 300$ and $10,000$ were used to examine the results for medium and large

data sets. For each of the simulation studies, $N = 1,000$ samples were generated. The SP-GLM-I and SP-GLM-H methods required specification of a link function for model fitting which we chose the log-link function for each model. For GLM model fitting, as well as the link function (log-link) we also had to specify the response distribution (Poisson).

The simulation results of $\hat{\boldsymbol{\beta}}$ with $\pi = 0.3$ are shown in Table 3.4. MEAN and BIAS of SP-GLM-I are relatively close to those of the GLM and SP-GLM-H methods. The smaller sample sizes mainly result in bigger biases than the larger sample sizes. The biggest biases for zero-inflated data are in the intercepts which are probably because of the zero inflation. Note that the SP-GLM-H method cannot handle sample size of $n = 10,000$.

The simulation results for the asymptotic and the Monte Carlo standard errors of SP-GLM-I show consistency between these two standard errors. For the case of $n = 300$, the AASEs and MCSEs of SP-GLM-I are similar to those of SP-GLM-H. On the other hand, while GLM provides accurate regression coefficient estimation, their asymptotic standard errors are not correct. These simulation results confirm the accuracy of our proposed asymptotic variance.

The MSEs decrease with increasing sample sizes (Table 3.4). For medium samples, the MSEs of SP-GLM-I are very close to the MSEs of SP-GLM-H. This can also indicate the accuracy of parameter estimation of SP-GLM-I. Although the MSEs of GLM are the lowest, they are inaccurate due to the errors in their asymptotic standard errors.

The ARSS of SP-GLM-I in Table 3.5 are very similar to ARSS of SP-GLM-H for medium samples and ARSS of GLM for both medium and large samples. This demonstrates that SP-GLM-I can fit the data well and is comparable with SP-GLM-H and GLM.

When it comes to the time for model fitting (ATIME), SP-GLM-I took more computational time than GLM (Table 3.5). However, SP-GLM-I was faster than SP-GLM-H. These figures show that the computational algorithm for our proposed SP-GLM-I is feasible and efficient

Table 3.4: Simulation results for zero-inflated Poisson response with zero proportion $\pi = 0.3$ and $n = 300$ and $10,000$.

		n = 300			n = 10,000		
Method		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
True $\boldsymbol{\beta}$		1.0000	-0.3000	0.5000	1.0000	-0.3000	0.5000
SP-GLM-I	MEAN	0.6300	-0.2912	0.5058	0.6421	-0.3002	0.5009
	BIAS	-0.3700	0.0088	0.0058	-0.3579	-0.0002	0.0009
	AASE	0.1530	0.1952	0.2053	0.0264	0.0334	0.0338
	MCSE	(0.1530)	(0.1997)	(0.2100)	(0.0266)	(0.0320)	(0.0342)
	MSE	0.1603	0.0382	0.0422	0.1288	0.0011	0.0011
SP-GLM-H	MEAN	0.6300	-0.2913	0.5059	-	-	-
	BIAS	-0.3700	0.0087	0.0059	-	-	-
	AASE	0.1513	0.1930	0.2030	-	-	-
	MCSE	(0.1531)	(0.1998)	(0.2101)	(-)	(-)	(-)
	MSE	0.1598	0.0373	0.0412	-	-	-
GLM	MEAN	0.6302	-0.2915	0.5057	0.6420	-0.3001	0.5009
	BIAS	-0.3698	0.0085	0.0057	-0.3580	-0.0001	0.0009
	AASE	0.1086	0.1394	0.1469	0.0186	0.0237	0.0240
	MCSE	(0.1525)	(0.1987)	(0.2092)	(0.0266)	(0.0319)	(0.0340)
	MSE	0.1485	0.0195	0.0216	0.1285	0.0006	0.0006

to implement.

Type I error

For testing Type I errors, the previous data were used. We set the null hypothesis to be $\beta_1 = -0.3$ and the alternative hypothesis is $\beta_1 \neq -0.3$. The Type I errors for each simulation with different sample sizes are shown in Table 3.6. Type I errors are tested at 5% significance level. While Type I errors for the SP-GLM-I method are acceptable and are comparable to

Table 3.5: ARSS and ATIME (in seconds) for the model fitting of zero-inflated Poisson response with zero proportion $\pi = 0.3$ and $n = 300$ and $10,000$.

		ARSS		ATIME	
Method	n	300	10,000	300	10,000
SP-GLM-I		1,216.9	41,323	0.022	0.412
SP-GLM-H		1,216.9	-	5.670	-
GLM		1,217.0	41,323	0.003	0.011

Type I errors for the SP-GLM-H method when $n = 300$, Type I errors based on WALD-AASE and LR for GLM are incorrect. For $n = 10,000$, the SP-GLM-H method is unable to produce results and Type I errors (WALD-AASE and LR) for the GLM method are incorrect, while the SP-GLM-I method can produce accurate Type I error rates that are close to the nominal rate (0.05). These are excellent type I errors, considering the SP-GLM model is misspecified for constant zero-inflated counts.

Table 3.6: Type I errors for β_1 for zero-inflated Poisson response with zero proportion $\pi = 0.3$ and $n = 300$ and $10,000$.

	n = 300			n = 10,000		
Method	WALD-AASE	WALD-MCSE	LR	WALD-AASE	WALD-MCSE	LR
SP-GLM-I	0.052	0.045	0.049	0.052	0.062	0.058
SP-GLM-H	0.056	0.045	0.051	-	-	-
GLM	0.174	0.048	0.174	0.158	0.064	0.158

The conclusion drawn from these simulations is that the SP-GLM-I method works well with zero-inflated Poisson data. The coefficient biases of SP-GLM-I are similar to the biases of SP-GLM-H and GLM. In addition, the SP-GLM-I and SP-GLM-H methods can

provide accurate asymptotic standard errors of regression coefficients, while GLM cannot. Thus the SP-GLM can offer more accurate inference than GLM when the specified response distribution is not a member of the exponential family and hence does not suit the model structure of GLM. Moreover, our SP-GLM-I method for fitting SP-GLM can handle larger data sets with less computational time.

3.5.3 Comparing different number of bins

In our method the number of bins m in SP-GLM-I must be specified and thus we wish to study the effect of varying m. In these simulation studies, we compared the effect of model fitting using different values of m for a continuous response.

The response data were generated from an exponential distribution with an inverse link function. We simulated $N = 1,000$ samples. Each sample had size $n = 500$ observations with the true coefficient values $\boldsymbol{\beta} = [1, 0.5, 0.8]^T$. The model fitting for GLM used the gamma distribution with an inverse link function, and for SP-GLM-I and SP-GLM-H we adopted the inverse link. Note that the exponential distribution is a special gamma distribution, so GLM should work well in this case.

The number of observations in each bin is assumed fixed and it is denoted as n_0. It is possible that the actual number of observations n_u in bin B_u is not equal to n_0 due to ties. Different n_0 were examined to compare the effect of m: n_0 was assigned to be 1, 5, 20, 25, 50 and 100 corresponding to m = 500, 100, 25, 20, 10 and 5 bins respectively.

The simulation results for the coefficient estimates with different m using the SP-GLM-I method are shown in Table 3.7. Table 3.8 shows the results of SP-GLM-I with $m = 500$ and simulation results from SP-GLM-H and GLM. Different m values provide similar biases and the MCSEs of $\hat{\boldsymbol{\beta}}$ for SP-GLM-I which are also similar to those for SP-GLM-H and GLM.

The AASEs and MSEs in SP-GLM-I generally decrease with increasing m. The gap between the asymptotic and the Monte Carlo standard errors is slightly higher for small m (5 and 10 bins) than for large m.

Table 3.9 shows the ARSS and the ATIME. The model fitting of SP-GLM-I with different m provide similar ARSS with slightly smaller for decreasing m. These results are also close to those of GLM and SP-GLM-H. These ARSS suggest our SP-GLM-I method can fit the data well with different m at a similar accuracy level as for the SP-GLM-H and GLM methods. SP-GLM-I used less model fitting time than SP-GLM-H, but more than GLM. The computational times of SP-GLM-I increase with increasing m. The time spent and the MSE values suggest the trade-off between the accuracy of the fitted model and the computational speed.

Type I error

The simulation results indicate that our proposed SP-GLM-I method performs well in parameter estimation regardless of the m value chosen. Our coefficient estimates are comparable to GLM and SP-GLM-H. Type I errors in Table 3.10 were generated based on H_0: $\beta_1 = 0.5$ and H_1: $\beta_1 \neq 0.5$ using the previous simulated data sets. Type I errors were provided based on the Wald test with asymptotic and Monte Carlo standard errors of $\hat{\beta}_1$, and likelihood ratio tests at 5% significance level. Type I errors from the Wald test with MCSEs of $\hat{\beta}_1$ in SP-GLM-I are all accurate. However, the results from Wald test with AASEs of $\hat{\beta}_1$ and likelihood ratio tests show that Type I error rates deviate from the nominal rate 5% for small m ($m = 5, 10$). So m needs to be large enough to show the features of the data. However, beyond some value of m, there is only a small impact on the model fitting for increasing m, while the computation time increases. Ruppert (2002) also conducted the testing on the impact of the number of knots

Table 3.7: Simulation results of SP-GLM-I with different number of bins (m) and $n = 500$.

m		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	True $\boldsymbol{\beta}$	1.0000	0.5000	0.8000
5	MEAN	1.0100	0.5083	0.8011
	BIAS	0.0100	0.0083	0.0011
	AASE	0.2034	0.3079	0.3177
	MCSE	(0.1743)	(0.2468)	(0.2645)
	MSE	0.0415	0.0948	0.1010
10	MEAN	1.0103	0.5087	0.8002
	BIAS	0.0103	0.0087	0.0002
	AASE	0.1929	0.2869	0.2949
	MCSE	(0.1735)	(0.2467)	(0.2636)
	MSE	0.0373	0.0824	0.0870
20	MEAN	1.0105	0.5088	0.7998
	BIAS	0.0105	0.0088	-0.0002
	AASE	0.1843	0.2709	0.2783
	MCSE	(0.1732)	(0.2469)	(0.2633)
	MSE	0.0341	0.0735	0.0774
25	MEAN	1.0105	0.5088	0.7998
	BIAS	0.0105	0.0088	-0.0002
	AASE	0.1820	0.2670	0.2743
	MCSE	(0.1732)	(0.2469)	(0.2632)
	MSE	0.0332	0.0714	0.0753
100	MEAN	1.0106	0.5087	0.7998
	BIAS	0.0106	0.0087	-0.0002
	AASE	0.1745	0.2549	0.2624
	MCSE	(0.1731)	(0.2471)	(0.2628)
	MSE	0.0305	0.0650	0.0689

Table 3.8: Simulation results with one observation in each bin ($n_0 = 1$) and $n = 500$.

m	Method		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	True $\boldsymbol{\beta}$		1.0000	0.5000	0.8000
500	SP-GLM-I	MEAN	1.0105	0.5090	0.7996
		BIAS	0.0105	0.0090	-0.0004
		AASE	0.1748	0.2539	0.2616
		MCSE	(0.1733)	(0.2472)	(0.2630)
		MSE	0.0307	0.0646	0.0684
-	SP-GLM-H	MEAN	1.0102	0.5090	0.8002
		BIAS	0.0102	0.0090	0.0002
		AASE	0.1696	0.2471	0.2548
		MCSE	(0.1735)	(0.2471)	(0.2632)
		MSE	0.0289	0.0611	0.0649
-	GLM	MEAN	1.0099	0.5092	0.8005
		BIAS	0.0099	0.0092	0.0005
		AASE	0.1713	0.2482	0.2559
		MCSE	(0.1730)	(0.2474)	(0.2627)
		MSE	0.0294	0.0617	0.0655

selection for a different problem in a penalized spline context and his results agree with what we found here.

These simulation results show the effectiveness of the SP-GLM-I method in handling small, medium and large data sets. It can handle different types of response data with fast convergence. In addition, we have also conducted simulations with larger sample sizes ($n \geq 10,000$) and more covariates ($q \geq 15$), the results will be reported in a paper we are preparing currently. Based on its performance in these simulations, we believe that the SP-GLM-I method is a good option for model fitting in practice.

Table 3.9: ARSS and ATIME (in seconds) for the model fitting of continuous response with different number of bins (m) and $n = 500$.

m	Method	ARSS	ATIME
5	SP-GLM-I	197.97	0.030
10	SP-GLM-I	198.00	0.039
20	SP-GLM-I	198.01	0.058
25	SP-GLM-I	198.02	0.064
100	SP-GLM-I	198.02	0.174
500	SP-GLM-I	198.03	0.938
-	SP-GLM-H	198.02	23.969
-	GLM	198.03	0.004

Table 3.10: Type I errors for β_1 for continuous response with different number of bins (m) and $n = 500$.

m	Method	WALD-AASE	WALD-MCSE	LR
5	SP-GLM-I	0.012	0.054	0.018
10	SP-GLM-I	0.024	0.052	0.020
20	SP-GLM-I	0.038	0.052	0.034
25	SP-GLM-I	0.042	0.052	0.038
100	SP-GLM-I	0.046	0.052	0.050
500	SP-GLM-I	0.048	0.054	0.040
-	SP-GLM-H	0.050	0.052	0.052
-	GLM	0.048	0.054	0.050

3.6 Conclusions

The SP-GLM relaxes the response distribution assumption required in GLMs by including a nonparametric component into the response distribution. This is an advantage in that fewer assumptions are required from the user. In model fitting, there is no need to specify

the response distribution, only the link function needs to be specified. Another strength of the SP-GLM is the ability to interpret the relationship between the expected mean and the coefficient parameters through the link function.

If the responses are generated from the parametric model in GLMs framework such as a Poisson distribution with log link and an exponential distribution with inverse link, the SP-GLM provides results comparable with the parametric model (GLM) which is expected to provide the best results. On the other hand, if the response distribution is unknown or beyond the set of standard distributions for GLM, e.g. the zero-inflated situation, the SP-GLM can provide better results than the parametric model (GLM). In this case, the GLM method shows discrepancy between the asymptotic standard errors and the Monte Carlo standard errors. As a result, the GLM method provides incorrect Type I error for the Wald test (with asymptotic standard error) and the likelihood ratio test.

In this chapter, a new computational algorithm for fitting the SP-GLM is proposed. The MI - Scoring algorithm is applied to iteratively estimate the reference density and the regression coefficients. The regression coefficients are estimated using the Fisher scoring algorithm. The MI algorithm is used to estimate the reference density. In addition, this MI algorithm does not require the inverse of the information or Hessian matrix for the parameter estimation, and thus reducing the computational burden in SP-GLM-H. Note that in our algorithm for SP-GLM the piecewise constant is used to approximate the reference density, however, kernel or spline methods can also be easily applied in the approximation of the reference density.

The simulation studies demonstrate that the asymptotic standard errors and the Monte Carlo standard errors of our method are close to each other, which provides evidence for the validity of the theoretical asymptotic standard error formula presented in Section 3.4. The results of the proposed method are comparable to the results of Huang's method for small and

medium samples. The proposed method is also effective and is capable to handle large data sets. Furthermore, the simulation studies also indicate that the model fitting results are not very sensitive to the choice of number of bins as long as the number of bins is large enough (e.g. $m \geq 20$ for $n = 500$) so that the features of the response data can be captured.

4

The semiparametric generalized linear model with canonical link

4.1 Introduction

GLMs are important tools of parametric regression analysis to explore the linear relationships between predictors and the response variable via a link function. However, to find the final

model with an appropriate combination of the response distribution and the link function may take time and can be difficult for an inexperienced modeller. So in many cases, modellers tend to use a canonical link function of a response variable distribution for model fitting. The canonical link function is a link function that has the canonical parameter θ_i equivalent to the linear predictor η_i. One of the obvious benefits of the canonical link is that the estimated response mean is, in most cases, assured to stay within the response variable's range (Breheny, 2013). Another benefit of using the canonical link is that the derivation of the maximum likelihood estimator is simplified (Breheny, 2013). Furthermore, some linear regression properties, e.g. the sum of the residuals (i.e. the difference between observed and predicted response values) equals zero, are guaranteed to hold (Breheny, 2013). In addition, the minimal sufficient statistic for the regression coefficients exists (Lindsey, 1997; Rodriguez, 2007).

In this chapter, we propose a novel regression model for a special case of the SP-GLM with unspecified canonical link function and a computational algorithm for fitting this model. The advantage of this model is that it is convenient and easy to use. In addition to its flexibility in handling the response distribution, users do not need to explicitly specify the link function as the data will automatically choose the canonical link function and fit it for the users. This is a kind of nonparametric link function. We do not need to see a mathematical expression for this canonical link function but can still visualize the canonical link function by plotting the linear predictor against the fitted mean. Examples of the visual displays to illustrate the implicit canonical link function will be shown in Section 4.5. Furthermore, this canonical link method can reduce the computational complexity of the SP-GLM, thus this method can handle larger data sets and with less computational time than the model fitting method for the SP-GLM discussed in Chapter 3.

In the following section, we explain the details of this SP-GLM with the canonical link function. The identifiability issue of the model is presented in Section 4.2.1. Then Section 4.3 describes the computational algorithm used to estimate the regression coefficients and the reference density. The asymptotic properties of this model are presented in Section 4.4. Visual displays of the canonical link function that is implicitly applied in model fitting, are shown in Section 4.5. Section 4.6 reports results from simulation studies. Finally, conclusions are set out in Section 4.7.

4.2 Semiparametric generalized linear model with canonical link

In this section, we summarize the main features of the SP-GLM that was explained in Section 2.3.1, then connect it with our proposed model with canonical link.

Observations y_i for $i = 1, \ldots, n$ corresponding to response variables Y_i for given $\mathbf{X}_i$ are assumed to be independent. The conditional distribution function $F(Y_i|\mathbf{X}_i)$ is a distribution in the exponential family with probability density function $f_i(y|\mathbf{X}_i) = f_0(y)\exp\{y\theta_i - b(\theta_i; f_0)\}$ where $b(\theta_i; f_0) = \log \int_{\mathcal{Y}} \exp\{y\theta_i\} f_0(y)\mathrm{d}y$. Denote the conditional mean and variance by $\mathrm{E}(Y_i|\mathbf{X}_i) = \mu_i$ and $\mathrm{Var}(Y_i|\mathbf{X}_i) = \sigma_i^2$ respectively. According to the exponential family properties, $\mu_i = b'(\theta_i; f_0)$ and $\sigma_i^2 = b''(\theta_i; f_0)$ for $a_i(\phi) = 1$.

In parametric GLMs, $f_0(\cdot)$ is a known function. However, as in Rathouz and Gao (2009) and Huang (2014), the SP-GLM assumes that $f_0(\cdot)$ is unknown and can be estimated from the data. Thus $f_0(\cdot)$ is nonparametric and $f_i(y|\mathbf{X}_i)$ can be seen as the exponential tilting form of the reference density f_0.

The linear predictor in GLMs is specified as $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a q-column vector of

regression coefficients. The linear predictor η_i is connected to the response mean μ_i through the link function $g(\cdot)$ that is $g(\mu_i) = \eta_i$.

θ_i is known as the canonical parameter or the natural parameter in GLMs. It is related to the response mean μ_i via the first derivative of $b(\theta_i; f_0)$ with respect to θ_i that is $\mu_i = b'(\theta_i; f_0)$. Then we define the inverse of the first derivative of $b(\cdot)$ to be a known monotonic function $z(\cdot)$, i.e. $z(\cdot) = b'^{-1}(\cdot)$. Then θ_i can be expressed as $\theta_i = z(\mu_i; f_0)$. If function $z(\cdot)$ is chosen to be the same function as the link function $g(\cdot)$, then $g(\cdot)$ is known as a canonical link function since this choice gives $\theta_i = \eta_i$. In GLMs, the canonical link function is extensively applied. For example, the well-known logistic regression employs the logit link with binary response distribution.

In the parametric distribution, the canonical link function and the canonical parameter can be explicitly specified. For instance, the normal distribution has $\theta_i = \mu_i$, so the canonical link function is the identity link. The canonical link function for the Poisson distribution is the log link with $\theta_i = \log(\mu_i)$. However, unlike the parametric GLMs where θ_i is clearly defined from the response distribution, the SP-GLM has to estimate θ_i that satisfies the mean constraint, $g^{-1}(\eta_i) = b'(\theta_i; f_0)$ as explained in Chapter 3.

In this chapter, we consider the special case of the SP-GLM when the canonical link function is chosen. This special link function gives $\theta_i = \eta_i$. The main advantage of this model is that it simplifies computations. It also reduces the complexity and time involved in estimating θ_i. Another advantage is that the modellers do not have to explicitly specify a link function since data itself will determine implicitly a canonical link function. This model is beneficial for inexperienced modellers. Even the experienced users can also enjoy the convenience of model building.

Under the SP-GLM with the canonical link, the probability density function can be written

as

$$f_i(y|\mathbf{X}_i) = f_0(y)\exp\{y\eta_i - b(\eta_i; f_0)\}, \tag{4.1}$$

where

$$b(\eta_i; f_0) = \log \int_{\mathcal{Y}} \exp\{y\eta_i\} f_0(y)\mathrm{d}y. \tag{4.2}$$

Then the mean and variance of y_i become $\mu_i = b'(\eta_i; f_0)$ and $\sigma_i^2 = b''(\eta_i; f_0)$ respectively.

4.2.1 Identifiability

As stated in Chapter 2, f_0 in the SP-GLM is not identifiable as the tilted version of f_0 can provide the same density function (2.7). This is because θ_i in (2.7) can be shifted by some constant. For this special SP-GLM with canonical link, to make the model identifiable, some restrictions are required on f_0 and $\boldsymbol{\beta}$. Basically, we require (i) the restriction on f_0: $\int_{\mathcal{Y}} f_0(y)dy = 1$ and (ii) the restriction on $\boldsymbol{\beta}$: $\beta_0 = 0$ where β_0 is the intercept. To clarify, let f_0^* be a tilted version of f_0 as

$$f_0^*(y) = f_0(y)\exp\{y\eta_0^* - b(\eta_0^*; f_0)\}, \tag{4.3}$$

where

$$b(\eta_0^*; f_0) = \log \int_{\mathcal{Y}} \exp\left\{y\eta_0^*\right\} f_0(y)\mathrm{d}y, \tag{4.4}$$

and η_0^* is any constant on $(-\infty, \infty)$. The identifiability issue exists if we also have

$$f_i(y|\mathbf{X}_i) = f_0^*(y)\exp\{y\eta_i^* - b^*(\eta_i^*; f_0^*)\}, \tag{4.5}$$

where

$$b^*(\eta_i^*; f_0^*) = \log \int_{\mathcal{Y}} \exp\left\{y\eta_i^*\right\} f_0^*(y)\mathrm{d}y. \tag{4.6}$$

Letting $\eta_i^* = \eta_i - \eta_0^*$ and we get $b^*(\eta_i^*; f_0^*) = b(\eta_i; f_0) - b(\eta_0^*; f_0)$ after substituting $f_0^*(y)$ from (4.3) in (4.6). This $f_0^*(y)$ can produce the same density function as $f_0(y)$ in (4.1). Substituting

$f_0^*(y)$ of (4.3) in (4.5) we get

$$f_i(y|\mathbf{X}_i) = f_0(y)\exp\left\{y(\eta_0^* + \eta_i^*) - \left(b(\eta_0^*; f_0) + b^*(\eta_i^*; f_0^*)\right)\right\}.$$

Thus η_i can be shifted by η_0^* and makes the model not identifiable. As $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$ and η_0^* is a constant, this η_0^* will only affect the intercept β_0. We need to impose the identifiability constraint that $\beta_0 = 0$, which gives $\eta_0^* = 0$. In many practical applications, the intercept is not of primary interest. So this should have no affect on model performance. With $\eta_0^* = 0$, $f_0^*(y)$ in (4.3) becomes $f_0^*(y) = f_0(y)\exp\{-b(\eta_0^*; f_0)\}$ where $b(\eta_0^*; f_0) = \log\int_{\mathcal{Y}} f_0(y)\mathrm{d}y$. Since $\int_{\mathcal{Y}} f_0(y)dy = 1$, then we get $b(\eta_0^*; f_0) = 0$ and it follows that $f_0^*(y) = f_0(y)$ as required.

The constraint on f_0 already exists and can be imposed by the Lagrange multipliers method. The constraint on β_0 can be imposed by centering the design matrix $\mathbf{X}$. That is the column mean of $\mathbf{X}$ be subtracted from the value of the corresponding column $\mathbf{X}$.

4.3 Computation algorithm

Let $\psi_u(\cdot)$ be non-negative basis functions and define its coefficients to be α_u. For computational purposes, the infinite dimensional parameter f_0 can be approximated using some non-negative basis functions:

$$f_0(y) = \sum_{u=1}^{m} \alpha_u \psi_u(y).$$

We have a restriction on α_u that $\alpha_u \geq 0$ to ensure the non-negativity of f_0.

The response data range is denoted as $\mathcal{H} = [y_{(1)}, y_{(n)}]$ where $y_{(1)} = \min\{y_1, \ldots, y_n\}$ and $y_{(n)} = \max\{y_1, \ldots, y_n\}$. Suppose B_u for $u = 1, \ldots, m$ is a partition of $\mathcal{H}$ where each partition is mutually exclusive and exhaustive, i.e. $\cup_{u=1}^{m} B_u = \mathcal{H}$ and if $u \neq v$, $B_u \cap B_v = \emptyset$. We define

$\psi_u(y)$ to be indicator basis functions

$$\psi_u(y) = \mathbf{1}_{B_u}(y) = \begin{cases} 1, & \text{if} \quad y \in B_u \\ 0, & \text{if} \quad y \notin B_u. \end{cases}$$

Then f_0 is regarded as a piecewise constant function. Suppose bin B_u has the width δ_u and define its probability mass as $p_u = \alpha_u \delta_u$. We have the constraints on p_u that (i) $p_u \geq 0$ for all u and (ii) $\sum_{u=1}^{m} p_u = 1$.

We define the number of response observations and the mid-point corresponding to bin B_u as n_u and c_u, respectively. Denote a m-vector for all p_u by $\boldsymbol{p}$ and let $\boldsymbol{\beta}$ be the vector for all β_j excluding the intercept. Then the log-likelihood function with this approximation to f_0 is

$$l(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} (y_i \eta_i - b(\eta_i; \boldsymbol{p})) + \sum_{u=1}^{m} n_u \log p_u, \tag{4.7}$$

where

$$b(\eta_i; \boldsymbol{p}) = \log \sum_{u=1}^{m} \exp\{c_u \eta_i\} p_u. \tag{4.8}$$

The mean and variance for observation y_i are

$$\begin{aligned} \mu_i &= b'(\eta_i; \boldsymbol{p}) = \sum_{u=1}^{m} c_u p_u \exp\{c_u \eta_i - b(\eta_i; \boldsymbol{p})\}, \\ \sigma_i^2 &= b''(\eta_i; \boldsymbol{p}) = \sum_{u=1}^{m} [c_u - \mu_i]^2 p_u \exp\{c_u \eta_i - b(\eta_i; \boldsymbol{p})\}. \end{aligned} \tag{4.9}$$

Note that the expression of $b(\eta_i; \boldsymbol{p})$ in (4.8) is a simplification obtained by applying c_u as a representative value of bin B_u. The exact result of approximating f_0 by a piecewise constant function yields

$$b(\eta_i; \boldsymbol{p}) = \log \sum_{u=1}^{m} \frac{p_u}{\eta_i} \left[\exp\{\max(B_u)\eta_i\} - \exp\{\min(B_u)\eta_i\} \right]. \tag{4.10}$$

We simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{p}$ using a special constrained maximum likelihood estimation method (named SP-GLM-CL). The equality constraint $\sum_{u=1}^{m} p_u = 1$ is imposed

through the method of Lagrange multipliers and the MI (Ma, 2010) algorithm is used to ensure that p_u satisfies the inequality constraints $p_u \geq 0$, for all u. Let the Lagrange multipliers be λ. Due to the equality constraint, the Lagrangian is given by

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{p}) &= l(\boldsymbol{\beta}, \boldsymbol{p}) - \lambda(1 - \sum_{u=1}^{m} p_u) \\ &= \sum_{i=1}^{n}(y_i\eta_i - b(\eta_i; \boldsymbol{p})) + \sum_{u=1}^{m} n_u \log p_u - \lambda(1 - \sum_{u=1}^{m} p_u). \end{aligned} \tag{4.11}$$

The KKT necessary conditions for the constrained maximum likelihood estimates are

$$\begin{aligned}&\frac{\partial \mathscr{L}}{\partial \beta_j} = 0 \qquad \text{for } j = 1, \ldots, (q-1), \\ &\frac{\partial \mathscr{L}}{\partial p_u} = 0 \text{ if } p_u > 0, \quad \frac{\partial \mathscr{L}}{\partial p_u} < 0 \text{ if } p_u = 0 \qquad \text{for } u = 1, \ldots, m, \\ &\frac{\partial \mathscr{L}}{\partial \lambda} = 0.\end{aligned}$$

We apply the Newton - MI algorithm (Ma et al., 2014) to iteratively estimate $\boldsymbol{\beta}$ and $\boldsymbol{p}$. Let $a^{(k)}$ be an estimate of any parameter a at iteration k. $\boldsymbol{\beta}$ is updated via the Newton algorithm

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \omega_1^{(k)}(\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}^{(k)}), \tag{4.12}$$

where $\mathbf{W} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ (see σ_i^2 in equation (4.9)), and $\mathbf{y}$ and $\boldsymbol{\mu}$ are vectors of y_i's and μ_i's respectively. A line search step size $\omega_1^{(k)} \in (0, 1]$ is applied in (4.12) to guarantee increasing log-likelihood when moving from $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+1)}$, namely $\mathscr{L}(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{p}^{(k)}) \geq \mathscr{L}(\boldsymbol{\beta}^{(k)}, \boldsymbol{p}^{(k)})$. This Newton algorithm is based on the score function for $\boldsymbol{\beta}$

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu})$$

and the second derivative of the log-likelihood with respect to $\boldsymbol{\beta}$

$$\frac{\partial^2 \mathscr{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X}.$$

Next, we apply the MI algorithm to update $p_u \geq 0$ by solving the KKT conditions on p_u that is

$$p_u \frac{\partial \mathscr{L}}{\partial p_u} = 0 \qquad \text{subject to} \quad p_u \geq 0, \qquad \text{for} \quad u = 1, \ldots, m,$$

where

$$\frac{\partial \mathscr{L}}{\partial p_u} = -\sum_{i=1}^{n} \exp\{c_u \eta_i \ - b(\eta_i; \boldsymbol{p})\} + \frac{n_u}{p_u} + \lambda. \tag{4.13}$$

We get $\lambda = 0$ by solving $\sum_{u=1}^{m} p_u \frac{\partial \mathscr{L}}{\partial p_u} = 0$. We employ the MI algorithm to estimate the iterative solution of p_u, even though p_u in (4.13) is non-linear. To ensure that p_u is non-negative, $p_u \frac{\partial \mathscr{L}}{\partial p_u}$ is manipulated to have non-negative values on both sides. Then, for all u, p_u is temporarily updated by

$$p_u^{(k+\frac{1}{2})} = \frac{n_u + \varepsilon}{\sum_{i=1}^{n} \exp\{c_u \eta_i^{(k)} - b(\eta_i^{(k)}; \boldsymbol{p}^{(k)})\} + \varepsilon}. \tag{4.14}$$

To avoid zero in the denominator, a small positive constant ε is added in (4.14). Note that the updated $\boldsymbol{p}$ is not affected by ε. The updated p_u obtained using equation (4.14) can guarantee non-negative values of p_u but the log-likelihood may not increase when moving from $\boldsymbol{p}^{(k)}$ to $\boldsymbol{p}^{(k+\frac{1}{2})}$. Thus to guarantee increasing log-likelihood, a line search step size $\omega_2^{(k)} \in (0, 1]$ is applied so that $\mathscr{L}(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{p}^{(k+1)}) \geq \mathscr{L}(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{p}^{(k)})$. Then $\boldsymbol{p}$ is updated by

$$\boldsymbol{p}^{(k+1)} = \boldsymbol{p}^{(k)} + \omega_2^{(k)} (\boldsymbol{p}^{(k+\frac{1}{2})} - \boldsymbol{p}^{(k)}). \tag{4.15}$$

If it converges, the solution will satisfy the KKT conditions on p_u. Equation (4.14) indicates that $p_u^{(k+\frac{1}{2})} \geq 0$ and thus (4.15) shows if $\boldsymbol{p}^{(k)} \geq \mathbf{0}$, then $\boldsymbol{p}^{(k+1)} \geq \mathbf{0}$. Here, the inequalities are interpreted element-wise. Note that $p_u^{(k+\frac{1}{2})} = 0$ only when $n_u = 0$. In this situation, it can be clearly seen from (4.13) that $\frac{\partial \mathscr{L}}{\partial p_u} < 0$.

We define the convergence criterion of the Newton - MI algorithm as being that the differences in absolute values of both $\boldsymbol{\beta}$ and $\boldsymbol{p}$ estimated in two consecutive iterations being less than 10^{-5}. We iteratively update $\boldsymbol{\beta}$ and $\boldsymbol{p}$ until the convergence criterion is achieved.

4.4 Asymptotic results

In this section we apply the same method for the asymptotic results in Chapter 3 to develop the asymptotic properties for the maximum likelihood estimates for parameters $\boldsymbol{\gamma} = \left(\boldsymbol{\beta}^T, \boldsymbol{p}^T\right)^T$. Let Γ be a domain of $\boldsymbol{\gamma}$. For the analysis of the asymptotic properties in this section we fix the dimension of $\boldsymbol{p}$, i.e. fix the number of bins m, therefore $\boldsymbol{\gamma}$ has length $(q-1)+m$. For fixed m, we assume that a density in the true data generating mechanism is the exponential tilting of a piecewise constant density. Suppose $\boldsymbol{\gamma}$ has the true parameter value $\boldsymbol{\gamma}^*$ and the maximum likelihood estimator $\hat{\boldsymbol{\gamma}}$. This $\hat{\boldsymbol{\gamma}}$ is an optimal solution that maximizes $l(\boldsymbol{\gamma})$ subject to $\sum_{u=1}^{m} p_u = 1$ and $p_u \geq 0$.

As mentioned in Chapter 3 that for the constrained problem, the usual asymptotic variance based on the unconstrained Fisher information matrix is incorrect since the active constraints affect the variance of the estimator. The active constraints presented in the parameter estimation algorithm have to be considered. Suppose the constraints are satisfied at a feasible point $\boldsymbol{\gamma}$. The equality constraint $\sum_{u=1}^{m} p_u = 1$ is always an active constraint, but the inequality constraint $p_u \geq 0$, $u = 1, \ldots, m$, will be an active constraint only if $p_u = 0$ and $\frac{\partial \mathscr{L}}{\partial p_u} < 0$. Assume the inequality constraints have m_1 active constraints, thus the total active constraints are $m_1 + 1$.

The Fisher information matrix for $\boldsymbol{\gamma}$ with active constraints requires matrix $U(\boldsymbol{\gamma})$ where the columns of matrix $U(\boldsymbol{\gamma})_{(q-1+m)\times(q+m-m_1-2)}$ form orthonormal bases for the null space of $G(\boldsymbol{\gamma})$, that is

$$G(\boldsymbol{\gamma})U(\boldsymbol{\gamma}) = \mathbf{0}_{(m_1+1)\times(q+m-m_1-2)} \quad \text{and} \quad U^T(\boldsymbol{\gamma})U(\boldsymbol{\gamma}) = \mathbf{I}_{(q+m-m_1-2)\times(q+m-m_1-2)}, \tag{4.16}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{0}$ is a zero matrix. We define $G(\boldsymbol{\gamma}) = [G_1(\boldsymbol{\beta}), G_2(\boldsymbol{p})]$ which has dimension $(m_1+1)\times(q-1+m)$. Let $G_1(\boldsymbol{\beta})$ be the first derivative of the active constraints

with respect to $\boldsymbol{\beta}$ which is a zero matrix with $(m_1 + 1) \times (q - 1)$ dimensions. The derivative of all active constraints with respect to $\boldsymbol{p}$ is denoted as $G_2(\boldsymbol{p})_{(m_1+1)\times m} = [G_3(\boldsymbol{p})^T, G_4(\boldsymbol{p})^T]^T$ where $G_3(\boldsymbol{p})$ is a m-row vector of 1s resulted from derivative of the equality constraint and matrix $G_4(\boldsymbol{p})$ contains the values of 0 and 1. For each row of $G_4(\boldsymbol{p})$, the value of 1 is contained only at the column u if $p_u = 0$ and $\frac{\partial \mathscr{L}}{\partial p_u} < 0$. Suppose $\boldsymbol{\tau}$ is a m-column vector of τ_u's where $\tau_u = 1$ if the corresponding inequality constraint is active, otherwise zero. Then $G_4(\boldsymbol{p})$ is defined by a matrix for which its rows are the rows of the identity matrix $\mathbf{I}_{m\times m}$ choosing if the corresponding $\tau_u = 1$, thus $G_4(\boldsymbol{p})$ has dimension $m_1 \times m$.

In order to develop the consistency and the asymptotic normality in Theorem 3, we require the following assumptions.

B1. For $i = 1, ..., n$, the random variables $\mathbf{X}_i$ are independent and identically distributed, and the distribution of $\mathbf{X}_i$ is independent of $\boldsymbol{\gamma}$.

B2. Γ is a compact subset of R^{q-1+m}.

B3. $\mathrm{E}_{\boldsymbol{\gamma}^*}[n^{-1}l(\boldsymbol{\gamma})]$ exists and has a unique maximum at $\boldsymbol{\gamma}^* \in \Gamma$.

B4. $l(\boldsymbol{\gamma})$ is twice differentiable in a neighborhood of $\boldsymbol{\gamma}^*$ and is continuous over Γ.

B5. The Fisher information matrix at $\boldsymbol{\gamma}^*$ exists and define as follows.

Let the $(q - 1) \times (q - 1)$ Fisher information matrix for $\boldsymbol{\beta}$ be

$$I_{\beta\beta} = -\mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right) = \mathbf{X}^T \mathbf{W} \mathbf{X}.$$

We define $I_{\boldsymbol{pp}}$ as the Fisher information matrix for $\boldsymbol{p}$ with $m \times m$ dimensions and its elements can be found from

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial p_u \partial p_v}\right) = -\sum_{i=1}^{n} \exp\{(c_u + c_v)\eta_i - 2b(\eta_i; \boldsymbol{p})\} + \frac{n_u}{p_u^2}\mathbf{1}_{u=v}.$$

If it is a diagonal element of $I_{\boldsymbol{pp}}$ ($u = v$), then $\mathbf{1}_{u=v}$ equals to 1, otherwise is zero. The $(q-1) \times m$ dimensional Fisher information matrix of β and $\boldsymbol{p}$ is defined as $I_{\boldsymbol{\beta p}}$ where its element corresponding to p_u is

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial p_u}\right) = \sum_{i=1}^{n} (c_u - \mu_i) \exp\{c_u \eta_i - b(\eta_i; \boldsymbol{p})\} \mathbf{X}_i.$$

Based on all of these elements, we have the Fisher information matrix for $\boldsymbol{\gamma}$ as

$$I(\boldsymbol{\gamma}) = -\mathrm{E}\left(\frac{\partial^2 l(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\right) = \begin{bmatrix} I_{\beta\beta} & I_{\beta \boldsymbol{p}} \\ I_{\boldsymbol{p}\beta} & I_{\boldsymbol{pp}} \end{bmatrix},$$

where $I_{\boldsymbol{p}\beta} = I_{\beta \boldsymbol{p}}^T$.

Theorem 3 *Suppose that assumptions B1 - B5 are satisfied, there are m_1 active inequality constraints and $U(\boldsymbol{\gamma}^*)$ is defined as in (4.16), then when $n \to \infty$, the constrained maximum likelihood estimator $\hat{\boldsymbol{\gamma}}$ is a consistent estimator for $\boldsymbol{\gamma}^*$ and*

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \xrightarrow{\mathcal{D}} N(0_{(q-1+m)\times 1}, F(\boldsymbol{\gamma}^*)^{-1}_{(q-1+m)\times(q-1+m)}),$$

where

$$F(\boldsymbol{\gamma})^{-1} = U(\boldsymbol{\gamma})(U^T(\boldsymbol{\gamma}) I(\boldsymbol{\gamma}) U(\boldsymbol{\gamma}))^{-1} U^T(\boldsymbol{\gamma}). \tag{4.17}$$

The proof is omitted here since it is simply modified from Moore et al. (2008) and Ma et al. (2017).

The accuracy of our asymptotic normality for the regression coefficients is examined via simulation studies in Section 4.6 by comparing the proposed asymptotic standard error results with the Monte Carlo standard error results as the Monte Carlo standard errors are considered to be an accurate estimate of the standard errors.

4.5 Display the canonical link function

In this section, we show that the implicit canonical link function used in our model can be identified via the plotting of the linear predictor $\hat{\eta}$ against the fitted mean $\hat{\mu}$. We use three data sets for the illustration. They are simulated binary, count and continuous response observations. We fit the canonical link function model to these generated data sets and then plot $\hat{\eta}$ against $\hat{\mu}$. In order to verify the accuracy of the estimated canonical links, we plot the true linear predictor $\eta^* = \mathbf{X}\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^*$ are the true coefficients, against the true mean μ^*, then compare the true canonical links with the estimated canonical links.

In the first simulated sample, we generated the response data from the binomial distribution with logit link for sample size of $n = 1,000$. The true coefficient was $\beta_1 = 2$. The covariate vector was $\mathbf{X}_i = [x_{i1}]$ where values for x_{i1} were randomly generated from the standard normal distribution.

For the second simulated sample, the response data were generated from the Poisson distribution with log link for sample of sizes $n = 1,000$. We set the true coefficient as $\beta_1 = 0.5$. The covariates were set to be the same as in the first sample.

The response data with sample size of $n = 1,000$ in the third simulated sample had the normal distribution with identity link. The true coefficient was $\beta_1 = 0.9$. We also used the same set of covariates from the first example.

Figures 4.1 - 4.3 present the plots of the true canonical links (left plots) and the estimated canonical links (right plots) for samples 1 to 3, respectively. These figures show that the canonical links are automatically chosen by these models. For all samples, the plots of the true canonical links indicate that the estimated canonical links are correct.

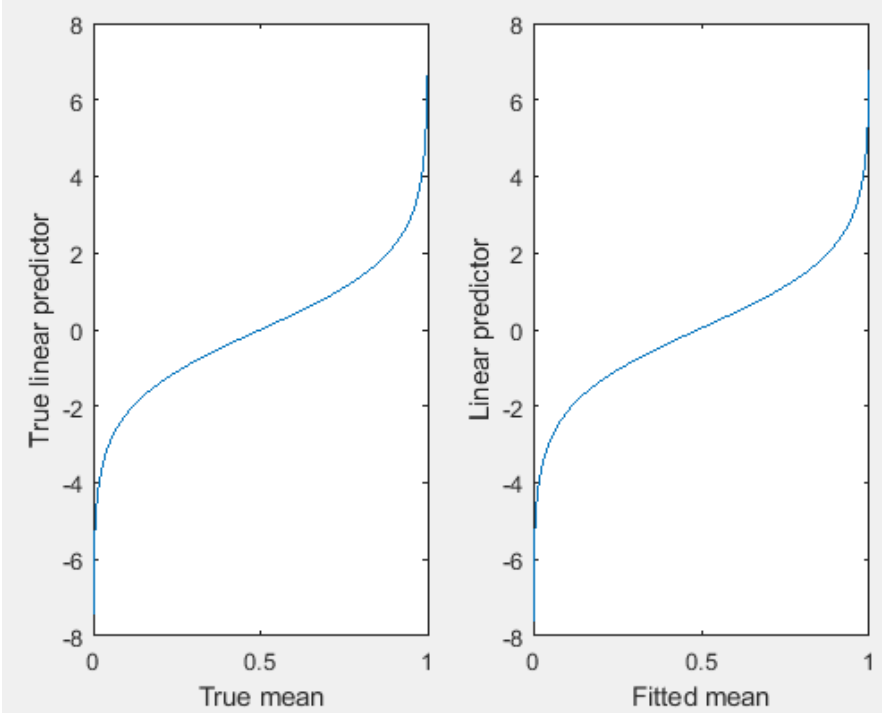

Figure 4.1: True mean (left) and fitted mean (right) curves of binomial response with logit link.

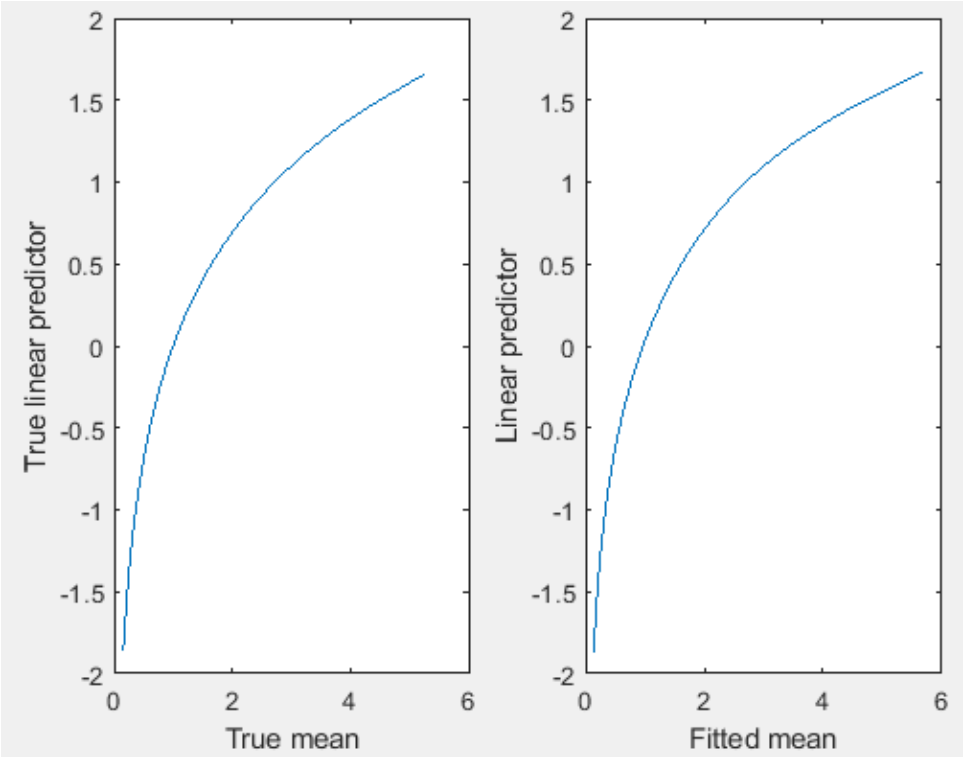

Figure 4.2: True mean (left) and fitted mean (right) curves of Poisson response with log link.

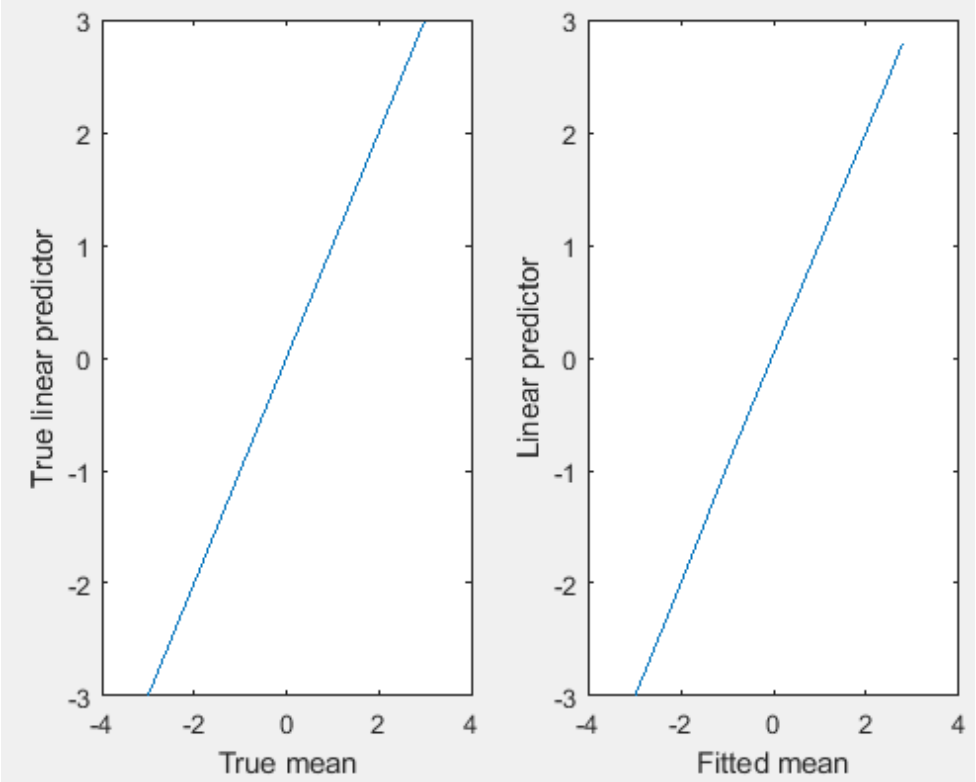

Figure 4.3: True mean (left) and fitted mean (right) curves of normal response with identity link.

4.6 Simulation studies

In this section, we conducted simulation studies with the aim of demonstrating the effectiveness of our implicit canonical link method. To indicate its accuracy, we compare the simulation results of the SP-GLM with unspecified canonical link function, named SP-GLM-CL, that we proposed in this chapter with the SP-GLM-I fitting method introduced in Chapter 3 and the classical GLM. We provide the MATLAB code for SP-GLM-CL and SP-GLM-I in Appendix A.2 and A.1, respectively. `fitglm()` MATLAB function is used for GLM.

In the following simulations, we generated three data sets. All response data were generated from the distributions within the GLM framework with their canonical links that were (i) the normal distribution with identity link, (ii) the binomial distribution with logit link and (iii) the Poisson distribution with log link. Thus we can explore the performance

of the SP-GLM-CL method in handling different types of response data (e.g. continuous, binary and count data) that are generated from different canonical link functions. In addition, different sample sizes are used in these simulated data sets, so that the effectiveness of the SP-GLM-CL method to fit the data with small, medium and large samples can be investigated.

In the first simulated data set, we also aim to explore the impact of different choice of the common number of observations n_0 in each bin on the model fitting results of SP-GLM-CL. The results of SP-GLM-CL with different n_0 are compared with those of SP-GLM-I and the simulation results from model fitting with $n_0 = 1$ are compared to those of GLM.

For each simulated data set, we conducted Monte Carlo simulations with $N = 1,000$ trials. To explore the goodness of fit of model fitting with SP-GLM-CL, the average residual sum of squares (ARSS) of SP-GLM-CL are calculated using formula (3.23) and compared with those of SP-GLM-I and GLM.

The accuracy of the $\boldsymbol{\beta}$ estimates of SP-GLM-CL is evaluated by comparing its $\hat{\boldsymbol{\beta}}$ with the $\hat{\boldsymbol{\beta}}$ of other specified link function methods (SP-GLM-I and GLM). We obtained $1,000$ estimated values and asymptotic standard errors of $\boldsymbol{\beta}$ from $1,000$ repeated samples of each Monte Carlo simulation for each method. Then for each simulation setting, we compute the average estimate (MEAN), bias (BIAS), average asymptotic standard error (AASE), Monte Carlo standard error (MCSE), and mean squared error (MSE) for the coefficient estimate of SP-GLM-CL by the formulas (3.17) - (3.21) and compare with those of SP-GLM-I and GLM. However, as specified in Section 4.2.1, the constraint on regression coefficients, $\beta_0 = 0$, is needed to make the semiparametric model identifiable, thus $\hat{\beta}_0$ was not estimated in the SP-GLM-CL model. Except for this SP-GLM-CL model, the intercepts were estimated but were omitted from the tables. Note that the AASE is used to calculate the MSE. The MCSE results shown in Tables in this section are put in brackets to avoid confusion with the AASE

results.

The accuracy of our asymptotic standard errors of $\hat{\boldsymbol{\beta}}$ as proposed in Section 4.4 is evaluated by comparing the AASE to the MCSE of SP-GLM-CL. The ASE of $\hat{\boldsymbol{\beta}}$ in SP-GLM-CL are calculated from the square root of the first $(q-1)$ diagonal values of the asymptotic covariance matrix in (4.17).

Speed of our computational algorithm for model fitting in SP-GLM-CL is explored by comparing the average computational time spend (ATIME) in seconds, computed by formula (3.24).

The accuracy of the coefficient estimate and inference is also examined through a hypothesis test. We generated the data under the null hypothesis that β_1 equals a specified value, then performed tests with Type I errors at 5% significance level for testing the null hypothesis by using (i) Wald test with AASE of $\hat{\beta}_1$ (WALD-AASE), (ii) Wald test with MCSE of $\hat{\beta}_1$ (WALD-MCSE) and (iii) likelihood ratio test for χ^2_1 (LR).

4.6.1 Impact of different number of observations in each bin (Normal distribution with identity link)

The SP-GLM-CL method uses the same technique to approximate f_0 as in the SP-GLM-I explained in Chapter 3. In both methods, the equal-frequency discretization technique is applied to set the indicator basis functions. We define n_0 to be the pre-specified number of observations in each bin. Note that this n_0 could be different from n_u which is the actual number of observations in bin B_u due to ties. In these simulation studies, we explore the impact of using different n_0 to our model fitting and the $\boldsymbol{\beta}$ estimates for a continuous response.

We generated $N = 1,000$ repeated samples of continuous response with sample sizes $n = 500$ from a normal distribution with identity link and the true coefficients were $[\beta_0, \beta_1, \beta_2]^T =$

$[1, -0.3, 0.5]^T$. The covariate vector was $\mathbf{X}_i = [1\ x_{i1}\ x_{i2}]$ where x_{i1} and x_{i2} were independently generated from the uniform distribution with the interval (0,1). We set $n_0 = 1, 5, 20, 25, 50$ and 100 and fitted SP-GLM-CL and SP-GLM-I corresponding to these settings. These assigned n_0 correspond to the number of bins $m = 500, 100, 25, 20, 10$ and 5 respectively. For GLM, there is no bin to specify and these models are comparable with models of SP-GLM-CL and SP-GLM-I with $n_0 = 1$.

Tables 4.1 - 4.2 show the simulation results of $(\hat{\beta}_1, \hat{\beta}_2)$ for normal response with different n_0. For SP-GLM-CL with increasing n_0, the BIAS increases while both standard errors decrease and the MSE decreases until $n_0 = 25$ then rises.

Table 4.1: Simulation results for one observation in each bin ($n_0 = 1$) with $n = 500$.

	SP-GLM-CL		SP-GLM-I		GLM	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
True $\boldsymbol{\beta}$	-0.3000	0.5000	-0.3000	0.5000	-0.3000	0.5000
MEAN	-0.3023	0.5090	-0.2989	0.5030	-0.2990	0.5032
BIAS	-0.0023	0.0090	0.0011	0.0030	0.0010	0.0032
AASE	0.1597	0.1589	0.1592	0.1564	0.1573	0.1545
MCSE	(0.1659)	(0.1587)	(0.1627)	(0.1523)	(0.1629)	(0.1524)
MSE	0.0255	0.0253	0.0254	0.0245	0.0247	0.0239

Comparing the results of SP-GLM-CL with SP-GLM-I, when we increase n_0, the BIAS in SP-GLM-CL grows much larger than in SP-GLM-I. The BIAS values of SP-GLM-I are consistent with different n_0. Except for $n_0 = 1, 5$, the difference between AASE and MCSE, the standard errors and the MSE are smaller in SP-GLM-CL than in SP-GLM-I. The results of these two methods are close when $n_0 = 1, 5$. For $n_0 = 1$, the results of SP-GLM-CL are close but the values are slightly larger than other methods. Thus if we want very small BIAS similar to other methods, we can choose a small number of n_0 e.g. $n_0 = 1, 5$. However, we

Table 4.2: Simulation results to compare different number of observations in each bin (n_0) with $n = 500$.

		SP-GLM-CL		SP-GLM-I	
n_0		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
	True $\boldsymbol{\beta}$	-0.3000	0.5000	-0.3000	0.5000
5	MEAN	-0.3005	0.5060	-0.2988	0.5030
	BIAS	-0.0005	0.0060	0.0012	0.0030
	AASE	0.1592	0.1585	0.1596	0.1567
	MCSE	(0.1649)	(0.1577)	(0.1627)	(0.1523)
	MSE	0.0253	0.0252	0.0255	0.0246
20	MEAN	-0.2779	0.4678	-0.2988	0.5031
	BIAS	0.0221	-0.0322	0.0012	0.0031
	AASE	0.1530	0.1522	0.1659	0.1630
	MCSE	(0.1530)	(0.1465)	(0.1626)	(0.1523)
	MSE	0.0239	0.0242	0.0275	0.0266
25	MEAN	-0.2697	0.4541	-0.2988	0.5030
	BIAS	0.0303	-0.0459	0.0012	0.0030
	AASE	0.1507	0.1499	0.1683	0.1653
	MCSE	(0.1485)	(0.1423)	(0.1626)	(0.1523)
	MSE	0.0236	0.0246	0.0283	0.0273
50	MEAN	-0.2332	0.3923	-0.2988	0.5030
	BIAS	0.0668	-0.1077	0.0012	0.0030
	AASE	0.1399	0.1388	0.1803	0.1769
	MCSE	(0.1295)	(0.1244)	(0.1626)	(0.1523)
	MSE	0.0240	0.0309	0.0325	0.0313
100	MEAN	-0.1843	0.3098	-0.2988	0.5030
	BIAS	0.1157	-0.1902	0.0012	0.0030
	AASE	0.1241	0.1226	0.2016	0.1976
	MCSE	(0.1035)	(0.1010)	(0.1625)	(0.1522)
	MSE	0.0288	0.0512	0.0406	0.0391

can also choose larger n_0 e.g. $n_0 = 20, 25$ that have larger BIAS, but smaller values of MSE.

The ARSS for model fitting are shown in Table 4.3. The ARSS are similar in all methods for $n_0 = 1, 5$. However, they slightly increased with increased n_0. ARSS for SP-GLM-CL are very close to SP-GLM-I except for $n_0 = 100$.

Table 4.3: ARSS for the model fitting of continuous response with different number of observations in each bin (n_0) for SP-GLM-CL and SP-GLM-I with $n = 500$.

n_0	SP-GLM-CL	SP-GLM-I	GLM
1	496.59	496.60	496.60
5	496.60	496.60	
20	496.64	496.60	
25	496.67	496.60	
50	496.93	496.60	
100	498.04	496.60	

Table 4.4 shows the time spent for model fitting. SP-GLM-CL was slower than GLM but much faster than SP-GLM-I. The computational time for SP-GLM-CL decreases with increasing n_0 until $n_0 = 20$. Then it is slightly fluctuates around the time required for $n_0 = 20$ which can be said that it is stable. These results suggest that we can choose any n_0 that is not too large compared to its sample sizes. However, since the trade-off between the goodness of fit and the computational time spend is presented, the medium size of n_0 that can still provide good fit with less time consumption is preferred.

Type I error

The accuracy of the parameter inferences of SP-GLM-CL with different n_0 is examined via hypothesis testing. We used the previous data sets that were generated based on the hypothesis that $\beta_1 = -0.3$ and did hypothesis testing H_0: $\beta_1 = -0.3$ and H_1: $\beta_1 \neq -0.3$ at

Table 4.4: ATIME for the model fitting of continuous response with different number of observations in each bin (n_0) for SP-GLM-CL and SP-GLM-I with $n = 500$.

n_0	SP-GLM-CL	SP-GLM-I	GLM
1	0.154	3.696	0.003
5	0.043	0.690	
20	0.012	0.231	
25	0.011	0.205	
50	0.008	0.131	
100	0.009	0.094	

the 5% significance level. Type I errors shown in Table 4.5 corresponding to the Wald test with asymptotic and Monte Carlo standard errors of $\hat{\beta}_1$, and the likelihood ratio test statistic. All methods provide appropriate error rates for $n_0 = 1$. For SP-GLM-I, the error rates are not appropriate for the WALD-AASE and LR for $n_0 = 100$. For SP-GLM-CL, the error rates are all approximately correct except the WALD-AASE for $n_0 = 100$, and WALD-MCSE and LR for $n_0 = 50, 100$. That is SP-GLM-CL for too large n_0 provide inaccurate parameter inference. So we need n_0 not to be too large.

In conclusion, n_0 in SP-GLM-CL can increase till some point yet still provide good parameter estimates and inferences while computational time decreases. However, from $n_0 = 20$ onwards, the time spent for model fitting is stable. The $\boldsymbol{\beta}$ estimates in SP-GLM-CL is more sensitive to very large n_0 than in SP-GLM-I. Compared to SP-GLM-I, the possible largest n_0 in SP-GLM-CL that can still provide accurate results is smaller than in SP-GLM-I. Hence, n_0 in SP-GLM-CL should not be too large relative to the sample sizes, so that the features of the data can still be captured and since the computational speed for large n_0 will not be improved from the medium n_0.

Table 4.5: Type I errors for β_1 for continuous response with different number of observations in each bin (n_0) for SP-GLM-CL and SP-GLM-I with $n = 500$.

n_0	Method	WALD-AASE	WALD-MCSE	LR
1	SP-GLM-CL	0.059	0.050	0.060
	SP-GLM-I	0.059	0.053	0.057
-	GLM	0.063	0.054	0.060
5	SP-GLM-CL	0.060	0.051	0.037
	SP-GLM-I	0.058	0.053	0.055
20	SP-GLM-CL	0.047	0.048	0.054
	SP-GLM-I	0.051	0.053	0.048
25	SP-GLM-CL	0.046	0.045	0.064
	SP-GLM-I	0.049	0.053	0.043
50	SP-GLM-CL	0.055	0.080	0.214
	SP-GLM-I	0.033	0.052	0.021
100	SP-GLM-CL	0.126	0.201	0.385
	SP-GLM-I	0.011	0.052	0.007

4.6.2 Binomial distribution with logit link

In the second simulated data set, the response variable was randomly generated from the binomial distribution with logit link which is its canonical link. Each simulated sample had sample size $n = 10,000$ and there were $N = 1,000$ repetitions. The covariates were generated the same way as in the previous simulated data sets. The true coefficients were $[\beta_0, \beta_1, \beta_2]^T = [-0.5, 0.3, 0.2]^T$.

In model fitting for GLM, the binomial distribution with logit link was specified. The logit link function was also used in SP-GLM-I. For SP-GLM-CL, neither the distribution of the response data nor the link function was required to be specified.

The simulation results for model fitting of data generated from a binomial distribution

with logit link shown in Table 4.6 indicate that SP-GLM-CL can produce accurate coefficient estimates and inferences, compared to SP-GLM-I and GLM. SP-GLM-CL can provide the same BIASes as other methods that had a specified canonical link (logit link) in their model fittings. The AASEs of $(\hat{\beta}_1, \hat{\beta}_2)$ for SP-GLM-CL are close to their corresponding MCSEs, and these standard errors are also the same as the standard errors provided by SP-GLM-I and GLM. Thus these results can confirm the accuracy of our asymptotic standard errors formula in (4.17).

Table 4.6: Simulation results for binary response with $n = 10,000$.

Method	SP-GLM-CL		SP-GLM-I		GLM	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
True $\boldsymbol{\beta}$	0.3000	0.2000	0.3000	0.2000	0.3000	0.2000
MEAN	0.3682	0.1858	0.3682	0.1858	0.3682	0.1858
BIAS	0.0682	-0.0142	0.0682	-0.0142	0.0682	-0.0142
AASE	0.7096	0.7266	0.7096	0.7266	0.7096	0.7266
MCSE	(0.7301)	(0.7537)	(0.7301)	(0.7537)	(0.7301)	(0.7537)
MSE	0.5081	0.5281	0.5081	0.5281	0.5081	0.5281

The ARSS of SP-GLM-CL in Table 4.7 also suggests that SP-GLM-CL can fit the data well at the same level as SP-GLM-I and GLM. Time taken in model fitting, on average in each repetition (seconds), is provided in Table 4.7. SP-GLM-CL spent slightly more computational time than GLM, but less than SP-GLM-I. Thus the performance of our computational algorithm is good, it can converge almost as fast as GLM.

Type I error

Next we perform the hypothesis testing, H_0: $\beta_1 = 0.3$ and H_1: $\beta_1 \neq 0.3$, at 5% significance level using the previous data set to examine the accuracy of the coefficient inferences of

Table 4.7: ARSS and ATIME for binary response with $n = 10,000$.

Method	SP-GLM-CL	SP-GLM-I	GLM
ARSS	23.820	23.820	23.820
ATIME	0.006	0.034	0.003

SP-GLM-CL. The previous data set was generated under the null hypothesis. Table 4.8 shows Type I error rates based on the likelihood ratio test, the Wald test with Monte Carlo and asymptotic standard errors of $\hat{\beta}_1$. Not surprisingly, Type I error rates for SP-GLM-CL are accurate. Note that since the AASE and MCSE values in all methods are the same, then the results from the Wald tests are also the same in all methods.

Table 4.8: Type I errors for β_1 for binary response with $n = 10,000$.

Method	SP-GLM-CL	SP-GLM-I	GLM
WALD-AASE	0.048	0.048	0.048
WALD-MCSE	0.051	0.051	0.051
LR	0.054	0.054	0.054

To conclude, our proposed SP-GLM-CL method works well with the binary response generated using a logit link. SP-GLM-CL correctly chosen the canonical link to fit the data. It can be used as an alternative regression method to the logistic regression as it can provide the same results with similar convergence speed as the model fitting in GLM using binomial distribution with logit link, while less information are required from users.

4.6.3 Poisson distribution with log link

For this simulated data set, the model fitting results of our SP-GLM-CL method were compared to those of other methods for count response. The response variable was randomly

generated from the Poisson distribution with its canonical link, that is the log link, for sample size $n = 100$ in $N = 1,000$ repeated samples. The covariates and the true coefficients were the same as in the second data set. In GLM, the Poisson distribution with log link was specified to fit the data, while only the log link function was specified for SP-GLM-I. For SP-GLM-CL, there was no need to specified the response distribution and link function.

Table 4.9 shows the Monte Carlo simulation results for these fitted models. SP-GLM-CL provides slightly higher BIAS for $\hat{\beta}_1$ but slightly less absolute BIAS value for $\hat{\beta}_2$ than other methods with log link. For SP-GLM-CL, the AASEs of $(\hat{\beta}_1, \hat{\beta}_2)$ are close to the MCSEs of $(\hat{\beta}_1, \hat{\beta}_2)$. Both standard errors in SP-GLM-CL are slightly higher than for the other methods with log link model. The AASEs that are matching with the MCSEs in SP-GLM-CL support the appropriateness of our proposed asymptotic variance. SP-GLM-CL has slightly higher MSEs than others with log link model.

Table 4.9: Simulation results for Poisson response with $n = 100$.

Method	SP-GLM-CL		SP-GLM-I		GLM	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
True $\boldsymbol{\beta}$	0.3000	0.2000	0.3000	0.2000	0.3000	0.2000
MEAN	0.3265	0.1970	0.3120	0.1868	0.3114	0.1866
BIAS	0.0265	-0.0030	0.0120	-0.0132	0.0114	-0.0134
AASE	0.4060	0.4001	0.3895	0.3841	0.3927	0.3874
MCSE	(0.3941)	(0.4065)	(0.3753)	(0.3850)	(0.3740)	(0.3842)
MSE	0.1655	0.1601	0.1519	0.1477	0.1544	0.1503

The ARSS for SP-GLM-CL in Table 4.10 is very close to that of the other methods with log link function. This result indicates that our SP-GLM-CL method works well for model fitting.

Table 4.10 also shows the time taken for model fitting. Our computational algorithm is very fast. Its computational time is comparable to GLM and it is much faster than SP-GLM-I.

Table 4.10: ARSS and ATIME for Poisson response with $n = 100$.

Method	SP-GLM-CL	SP-GLM-I	GLM
ARSS	75.586	75.582	75.585
ATIME	0.004	0.014	0.004

Type I errors

To examine the accuracy of the statistical inference of our SP-GLM-CL method, we perform the following hypothesis test. We tested the previous data under H_0: $\beta_1 = 0.3$ and H_1: $\beta_1 \neq 0.3$ at 5% significance level. Type I errors computed according to the likelihood ratio test and the Wald test using the asymptotic and the Monte Carlo standard errors of $\hat{\beta}_1$ in Table 4.11 suggest reasonable parameter inferences for the SP-GLM-CL method since its error rates are close to the nominal rate 0.05.

Table 4.11: Type I errors for β_1 for Poisson response with $n = 100$.

Method	SP-GLM-CL	SP-GLM-I	GLM
WALD-AASE	0.041	0.042	0.038
WALD-MCSE	0.063	0.048	0.047
LR	0.045	0.041	0.041

To summarize, our proposed SP-GLM-CL method performs well for model fitting of the count response data generated from the Poisson distribution with log link. The results of the SP-GLM-CL method are comparable with those of the other methods with specified canonical link function. These results support the effectiveness of our model fitting method

and the accuracy of our regression coefficient estimations and inferences. Furthermore, our proposed asymptotic standard errors are comparable with the Monte Carlo standard errors. This supports the suitability of our proposed asymptotic variance. In addition, our method uses much less computational time than SP-GLM-I, it is as fast as the GLM method.

4.7 Conclusions

In this chapter, the novel model and its algorithm for a special case of the SP-GLM when the canonical link function is applied, are proposed. In addition to the relaxation of the response distribution by including the unspecified reference density, this proposed canonical link model also relaxes the requirement of the specified link function in the SP-GLM by implicitly applying the canonical link function in the model. Thus in addition to other advantages of the SP-GLM, this canonical link setting can make the model fitting algorithm converge faster than the algorithm fitting method proposed in Chapter 3 since the computational complexity of the SP-GLM is reduced. Moreover, the estimated canonical link can be visualized via the plot of the linear predictor and the fitted mean.

In model fitting, we used the Newton - MI algorithm to simultaneously estimate the regression coefficients and the reference density. Even though we applied the indicator basis functions to approximate the reference density, other nonparametric approximation methods such as spline or kernel can also be applied in this algorithm. The MI algorithm that is used to estimate the reference density and also impose the non-negativity constraints on the reference density, has less computational cost than other traditional iterative methods (e.g. the Fisher scoring method) since this MI algorithm does not require the inverse of the Fisher information matrix which possibly has large dimension.

The simulation results showed that our proposed model fitting method for the SP-GLM

with unspecified canonical link function works well with data that are generated from parametric distributions with its canonical link such as the binomial distribution with logit link, Poisson distribution with log link and normal distribution with identity link, for small, medium and large sample sizes. The coefficient estimates and inferences are accurate and are comparable with the GLM and SP-GLM-I methods that are fitted with the correctly specified canonical link functions. Moreover, the accuracy of the asymptotic standard errors discussed in Section 4.4 is supported by our simulation results. For continuous response data, the $\boldsymbol{\beta}$ estimates and inferences are not very sensitive to the choice of n_0 as long as n_0 is not too large compared to its sample size. Furthermore, the proposed SP-GLM-CL method can handle sample sizes in the millions using a small number of bins. In addition, the computational speed for model fitting algorithm is very fast and it is comparable with GLM. Thus our proposed model fitting method is a good choice for regression analysis due to its simplicity and computational accuracy and efficiency.

5

Application to real data sets

In this chapter, we aim to demonstrate the performance of the SP-GLM-I and SP-GLM-CL methods proposed in Chapters 3 and 4 with real data sets. This chapter provides comparative results of applying the SP-GLM-I and SP-GLM-CL methods to the following real data sets (1) Vehicle insurance data (De Jong & Heller, 2008), (2) PhD students' research productivity data (Long, 1990) and (3) CD4 count data (Wade & Ades, 1994). For the first application, the model fitting results of the SP-GLM-I, SP-GLM-CL and GLM methods are compared by

using a vehicle insurance policies data set. This data set is typical of a real insurance data set which is relatively large (number of observations = 67,856; number of variables = 10). The SP-GLM-H method cannot handle this volume of data. In the second application, we compare the regression analysis of the PhD biochemists' article data set (number of observations = 915; number of variables = 6) using the SP-GLM-I, SP-GLM-CL and GLM methods. For the last application, a CD4 (number of lymphocytes per cubic millimeter of blood) count data (number of observations = 609; number of variables = 2) is used as a test data set to compare the model fitting results between the proposed SP-GLM-CL and the GLM methods.

The MATLAB code for the SP-GLM-I and SP-GLM-CL model fitting are provided in Appendix A.1 and A.2, respectively. For the GLM, the `fitglm()` MATLAB function is used to fit the model.

5.1 Vehicle insurance data

In this example, we demonstrate application of the SP-GLM-I and SP-GLM-CL methods to the one-year vehicle insurance policies data given by De Jong and Heller (2008). This data set contains 67,856 vehicle insurance policies that were taken out in 2004 or 2005. The variables in this data set are described in Table 5.1. For insurance companies, it is important to investigate the risk factors that influence policyholders to make claims, to predict the number of claims and claim size. The insurance companies use these data to determine the insurance premiums.

De Jong and Heller (2008) used this data set to demonstrate the risk factors associated with making claims (`clm`) by vehicle insurance policyholders. In this example, we are interested in the regression analysis with the response variable `numclm`, which is the number of claims for each vehicle insurance policy. The frequencies for the count response, `numclm`, are shown

Table 5.1: Variables from vehicle insurance policies.

Variables	Description
`clm`	Claim occurrence (1 = have at least one claim, 0 = have no claim).
`numclm`	Number of claims.
`clmcst`	Claim amount.
`exposure`	Exposure, ranging from 0 to 1.
`veh_val`	Vehicle value in $10,000s.
`veh_body`	Vehicle body category, coded as BUS, CONVT (convertible), COUPE, HBACK (hatchback), HDTOP (hardtop), MCARA (motorized caravan), MIBUS (minibus), PANVN (panel van), RDSTR (roadster), SEDAN, STNWG (station wagon), TRUCK, UTE (utility) where SEDAN is the reference category.
`veh_age`	Vehicle's age.
`gender`	Driver's gender, coded as M (Male), F (Female) where M is the reference category.
`agecat`	Driver's age category, coded as 1 to 6 where 1 is the youngest and is the reference category.
`area`	Driver's residential area, coded as A to F where A is the reference category.

in Table 5.2. The possible risk factors for `numclm` are vehicle value (`veh_val`), vehicle body type (`veh_body`), vehicle's age (`veh_age`), driver's gender (`gender`), driver's age (`agecat`) and driver's residential area (`area`).

We fit a Poisson regression (GLM), SP-GLM-I and also SP-GLM-CL to this data set. Note that for GLM, both the Poisson response distribution and log link need to be specified, while for SP-GLM-I only the log link is required. We do not need to specify the response distribution nor the link function for SP-GLM-CL.

The amount of exposure during the year is denoted as `"exposure"`. Exposure takes

Table 5.2: The frequency table of the number of claims.

Value	Count	Percent
0	63,232	93.19%
1	4,333	6.39%
2	271	0.40%
3	18	0.03%
4	2	0.00%

a value between 0 and 1 where 1 indicates full exposure. Each policy may have different exposure to risk, thus the adjustment for differing observation periods is required. That is the number of claims for each policy is adjusted to be per unit time (`numclm/exposure`). Since we fit the model using log link function for the SP-GLM-I and GLM methods, we adjust for exposure to risk of each policy by including log of `exposure` as the offset in the model. We assume the same offset for the SP-GLM-CL method to be able to compare its result with the results of SP-GLM-I and GLM methods. This pre-defined offset function for SP-GLM-CL is appropriate since in this example, the SP-GLM-CL model is automatically selected the log link in its model fitting as indicated by plotting the linear predictor and fitted mean shown in Figure 5.1.

The results of the SP-GLM-I and SP-GLM-CL methods are compared with the results obtained from GLM. These proposed models are assessed by using the histogram and Probability - Probability (P-P) plots of PIT (nonrandomized version of the probability integral transform (Czado, Gneiting, & Held, 2009)). The PIT is used to assess the model fit, to indicate appropriateness of an implicit distribution to fit the SP-GLM and to check an underlying distribution assumption in GLM (Czado et al., 2009; Fung & Huang, 2016). We follow the PIT formula of Czado et al. (2009). If the model fitting uses a correct response distribution

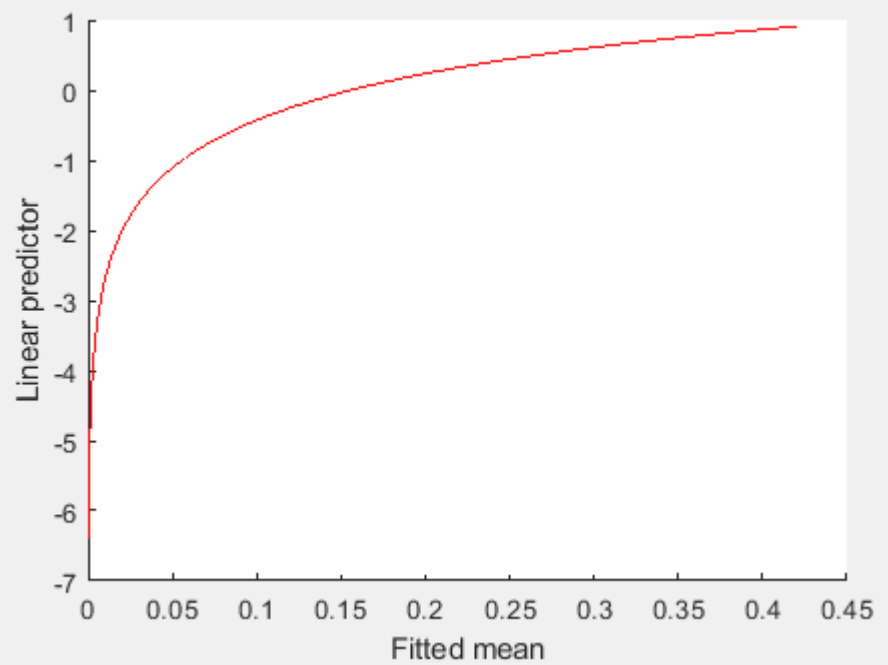

Figure 5.1: The fitted mean curve of SP-GLM-CL model for the vehicle insurance data set.

and the model fit is good, the PITs should follow a standard uniform distribution. That is the PIT histogram should be uniform and the P-P plot should follow the 45° line. The P-P plot compares the cumulative distribution of PIT with the standard uniform cumulative distribution. Departure from uniformity suggests the possibility of model insufficiency (Czado et al., 2009). An example of the use of the PIT histogram as a diagnostic check for an unspecified distribution of the SP-GLM can be found in Fung and Huang (2016).

The final model results for SP-GLM-I, SP-GLM-CL and GLM are shown in Table 5.3. For all three methods, `veh_body`, `veh_age` and `agecat` are the risk factors that are significant at 5% of significance level to explain the expected number of claims made by the policyholders. Both parameter estimates and standard errors for SP-GLM-I and SP-GLM-CL are close to those for GLM.

Figures 5.2 - 5.4 are the PIT histograms of SP-GLM-I, SP-GLM-CL and GLM, respectively. All plots appear to be uniform. In addition, the P-P plots of PITs as shown in Figure

Table 5.3: The final model results of the vehicle insurance data set.

		SP-GLM-I		SP-GLM-CL		GLM	
Parameter		Estimate	ASE	Estimate	ASE	Estimate	ASE
Intercept		-1.4038	0.0656	-	-	-1.4149	0.0602
veh_body	BUS	0.9000	0.3862	0.8679	0.3037	0.9191	0.3175
veh_body	CONVT	-0.5892	0.5982	-0.5826	0.5729	-0.5976	0.5780
veh_body	COUPE	0.4265	0.1311	0.4086	0.1159	0.4206	0.1185
veh_body	HBACK	-0.0563	0.0402	-0.0586	0.0368	-0.0603	0.0374
veh_body	HDTOP	0.0963	0.0980	0.0968	0.0880	0.1017	0.0897
veh_body	MCARA	0.5817	0.2944	0.5540	0.2525	0.5776	0.2596
veh_body	MIBUS	-0.0647	0.1629	-0.0474	0.1498	-0.0497	0.1519
veh_body	PANVN	0.0504	0.1366	0.0576	0.1217	0.0639	0.1241
veh_body	RDSTR	0.3524	0.6647	0.3884	0.5639	0.4024	0.5782
veh_body	STNWG	0.0331	0.0412	0.0321	0.0375	0.0333	0.0381
veh_body	TRUCK	-0.0505	0.0993	-0.0376	0.0900	-0.0374	0.0915
veh_body	UTE	-0.2037	0.0699	-0.1913	0.0646	-0.1971	0.0656
veh_age		-0.0625	0.0147	-0.0649	0.0134	-0.0665	0.0136
age_cat	2	-0.1852	0.0592	-0.1624	0.0531	-0.1682	0.0542
age_cat	3	-0.2413	0.0577	-0.2214	0.0518	-0.2289	0.0529
age_cat	4	-0.2719	0.0575	-0.2501	0.0517	-0.2587	0.0527
age_cat	5	-0.4940	0.0638	-0.4638	0.0578	-0.4779	0.0590
age_cat	6	-0.4857	0.0727	-0.4519	0.0662	-0.4656	0.0675

5.5 for SP-GLM-I (dashed-dotted green line), SP-GLM-CL (dashed red line) and GLM (solid blue line) models all follow the 45° line. So these indicate that the model fit of these methods are good. This also means the SP-GLM-I and SP-GLM-CL methods use appropriate response distributions in their model fittings.

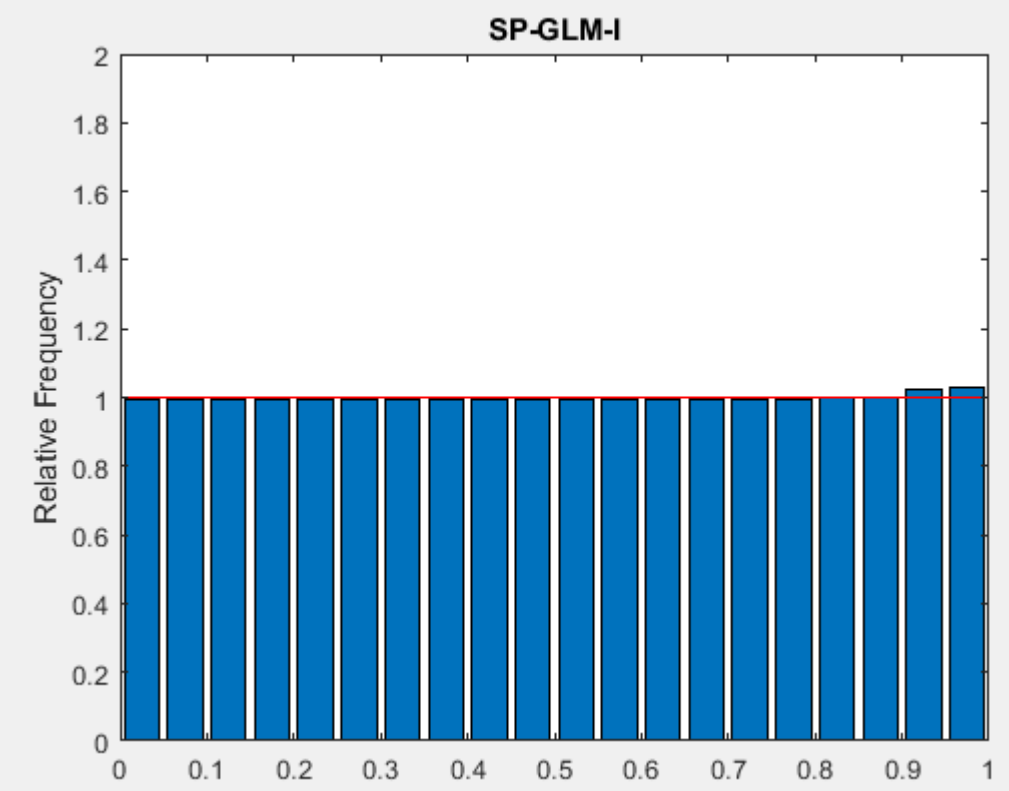

Figure 5.2: PIT histogram of SP-GLM-I for the vehicle insurance data set.

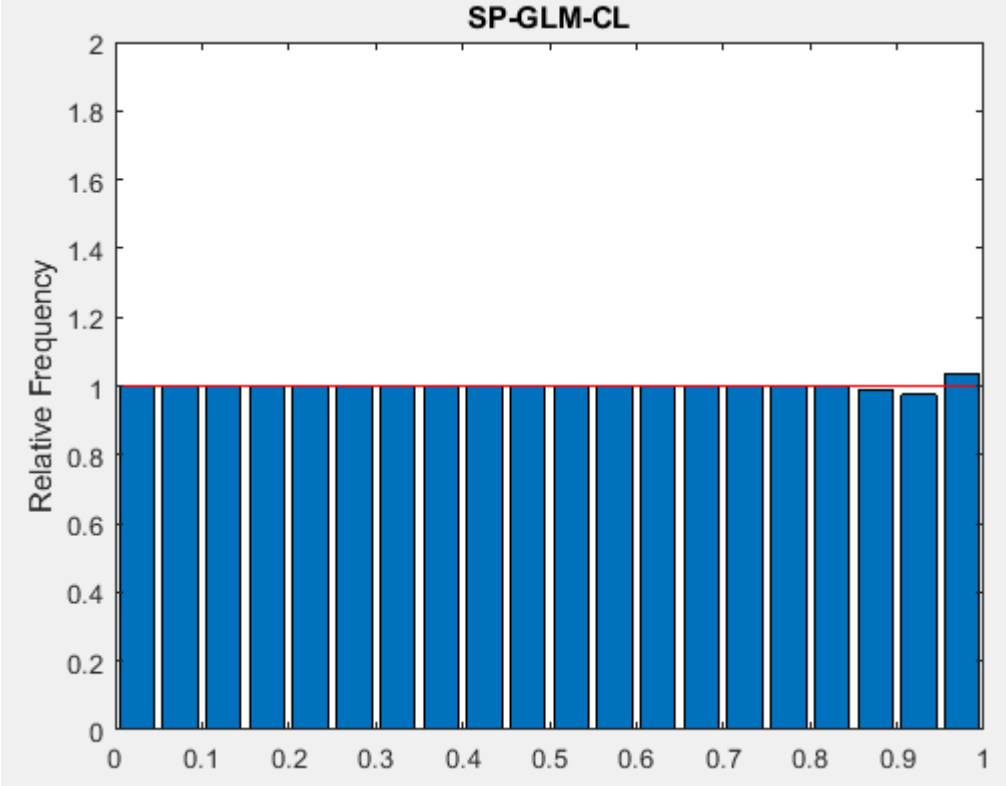

Figure 5.3: PIT histogram of SP-GLM-CL for the vehicle insurance data set.

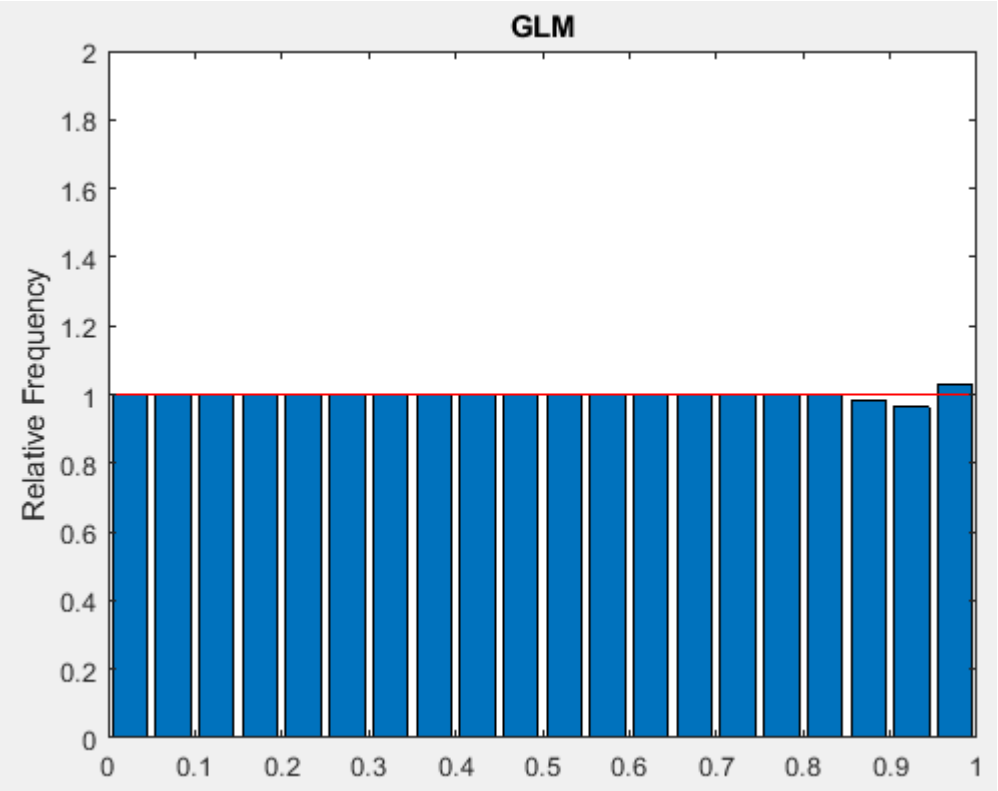

Figure 5.4: PIT histogram of GLM for the vehicle insurance data set.

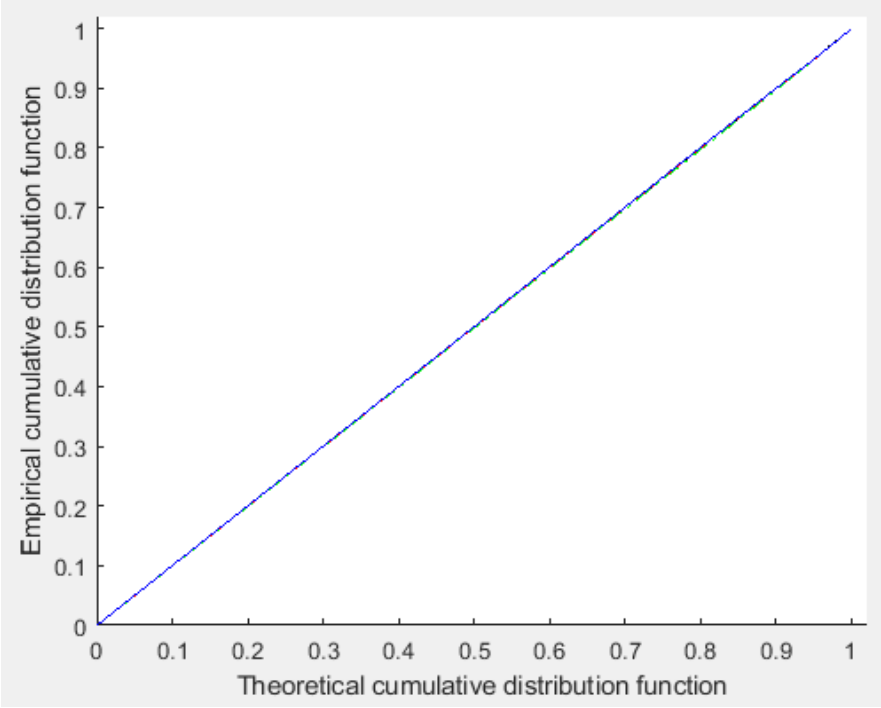

Figure 5.5: The P-P plots of PITs for SP-GLM-I (dashed-dotted green), SP-GLM-CL (dashed red) and GLM (solid blue) with the vehicle insurance data set.

To conclude, this example shows the effective performance of the SP-GLM-I and SP-GLM-CL methods in application to a real data set. They can handle multiple and different types of covariates in very large samples where the offset is also included in the model.

5.2 Research productivity of PhD graduates data set

Long (1990) studied the effects of gender on the productivity of PhD students in biochemistry. In this example, we perform a regression analysis on the mean number of articles produced by biochemistry PhD students in the last 3 years of their PhD (`art`). The frequency and the histogram of `art` are shown in Figure 5.6 and Table 5.4, respectively. The suggested variables by Long (1990) that affect the number of PhD articles are gender (`fem`: female [reference group], male), marital status (`mar`: single [reference group], married), number of children aged under 6 years old (`kid5`), PhD department's prestige (`phd`) and the number of articles by PhD mentors in last 3 years (`ment`) (Long, 1997). This data set is the `bioChemists` data given in the R package `pscl` (Jackman, 2017).

We apply the SP-GLM-I and SP-GLM-CL methods as well as the Poisson with log link model (GLM) to fit this data set. The performance of each method and the diagnostics for the underlying distribution assumption is checked by the PIT.

The final model results for all three methods are given in Table 5.5. All variables except the prestige of department (`phd`) are significant predictors for the expected number of publications by PhD biochemists. The absolute coefficient estimate values for SP-GLM-I are higher than the results of SP-GLM-CL, but are lower than those of GLM (except for `ment`). Even though the estimated coefficient parameters of each method are not the same, the predictors of all models show similar effect (e.g. increase / decrease) to the average number of articles. The coefficient standard errors of SP-GLM-I are the highest compared

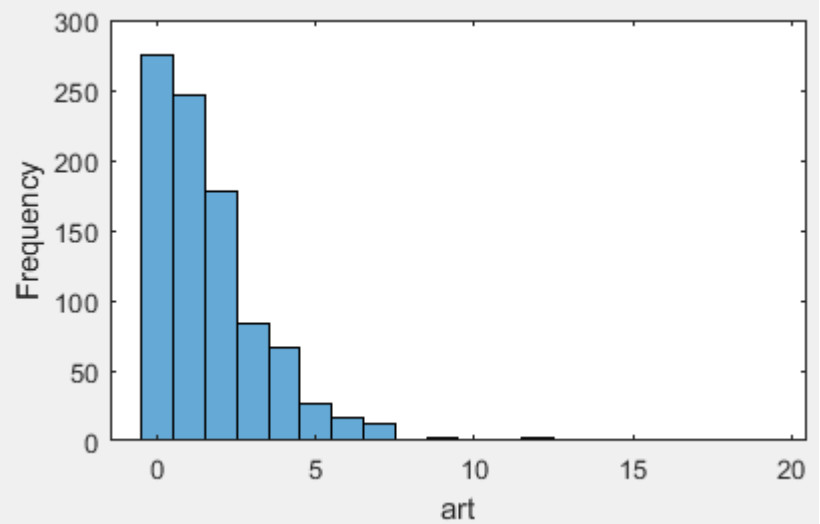

Figure 5.6: Histogram of `art`.

Table 5.4: The frequency table of the number of articles.

Value	Count	Percent
0	275	30.05 %
1	246	26.89 %
2	178	19.45 %
3	84	9.18 %
4	67	7.32 %
5	27	2.95 %
6	17	1.86 %
7	12	1.31 %
8	1	0.11 %
9	2	0.22 %
10	1	0.11 %
11	1	0.11 %
12	2	0.22 %
16	1	0.11 %
19	1	0.11 %

Table 5.5: The final model results of the PhD articles data set.

	SP-GLM-I		SP-GLM-CL		GLM	
Parameter	Estimate	ASE	Estimate	ASE	Estimate	ASE
intercept	0.091	0.077	-	-	0.120	0.054
`fem`	0.199	0.080	0.124	0.041	0.225	0.055
`mar`	0.140	0.090	0.083	0.045	0.152	0.061
`kid5`	-0.169	0.057	-0.106	0.030	-0.185	0.040
`ment`	0.029	0.004	0.011	0.002	0.026	0.002

to those of SP-GLM-CL and GLM. Note that since the link function in the SP-GLM-CL method is implicitly determined, the interpretations of its coefficient estimates are different from other methods where the link function is specifically specified.

The PIT histograms of SP-GLM-I and SP-GLM-CL in Figures 5.7 - 5.8 are close to uniform, while the histogram of the PITs for GLM (Figure 5.9) is U-shaped. Figure 5.10 shows the P-P plots of PITs for all methods. The P-P plots for both SP-GLM-I (dashed-dotted green line), SP-GLM-CL (dashed red line) are very close to the comparison line (dotted black line), but the plot for GLM (solid blue line) deviate from the 45° line. These plots indicate good model fit for SP-GLM-I and SP-GLM-CL, while for the GLM method they indicate a lack of model fit. The response distributions are appropriate for SP-GLM-I and SP-GLM-CL, but there is an inappropriate underlying distribution for GLM and thus its standard errors and inferences may be biased.

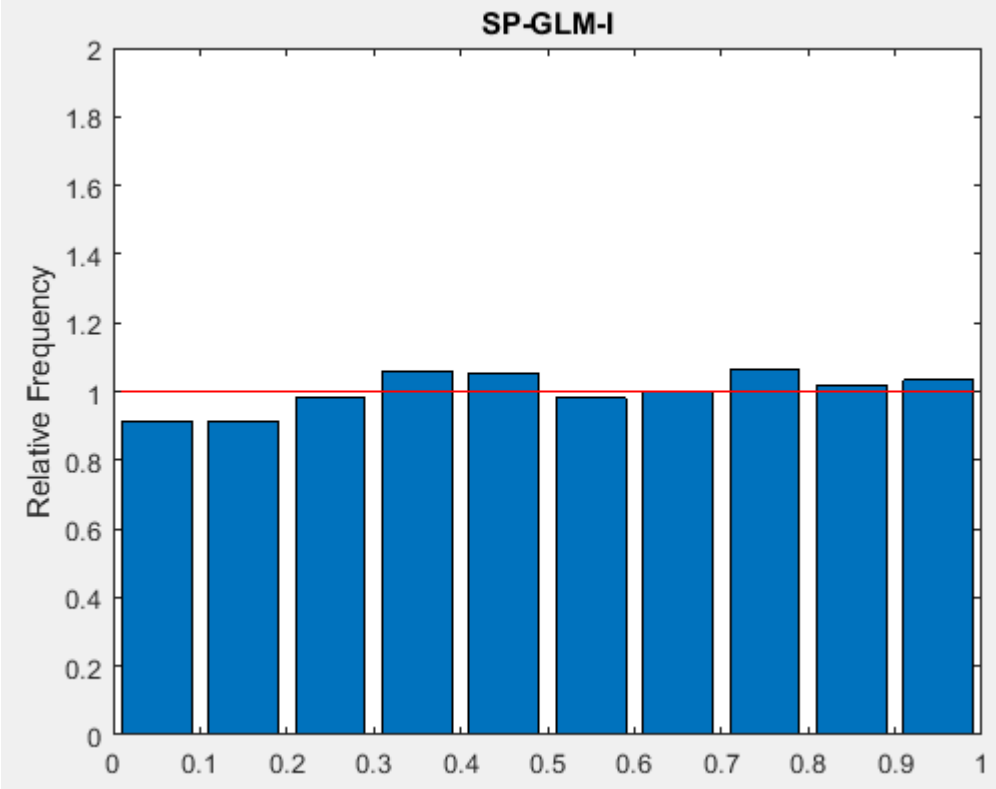

Figure 5.7: PIT histogram of SP-GLM-I for the PhD articles data set.

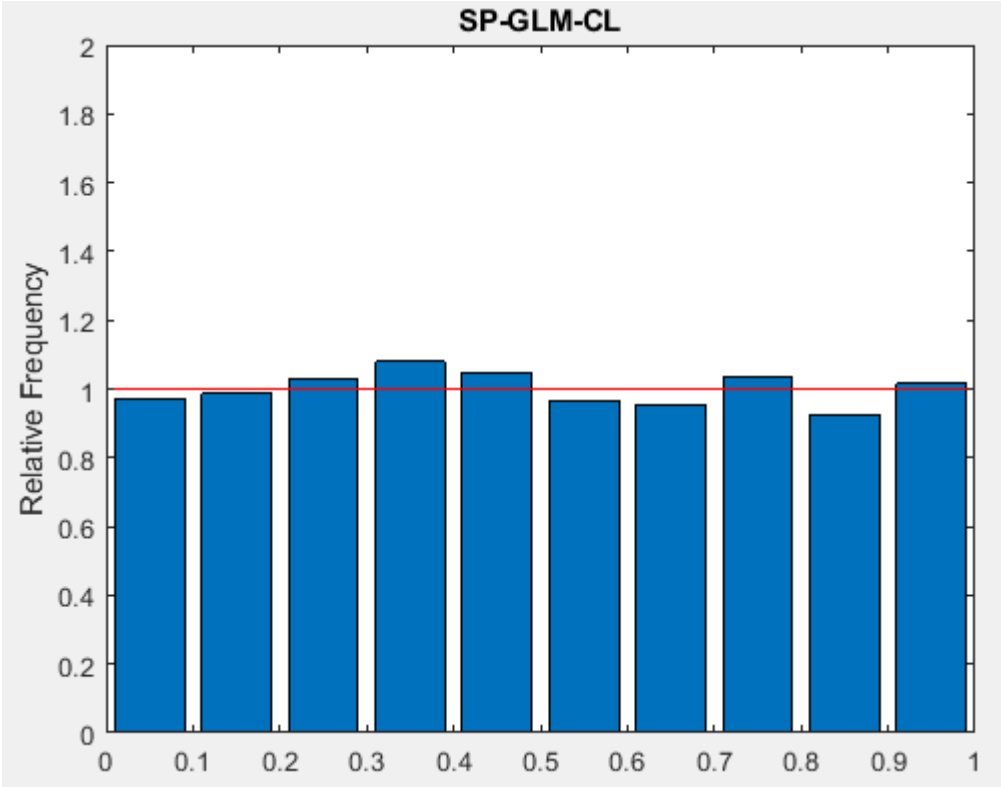

Figure 5.8: PIT histogram of SP-GLM-CL for the PhD articles data set.

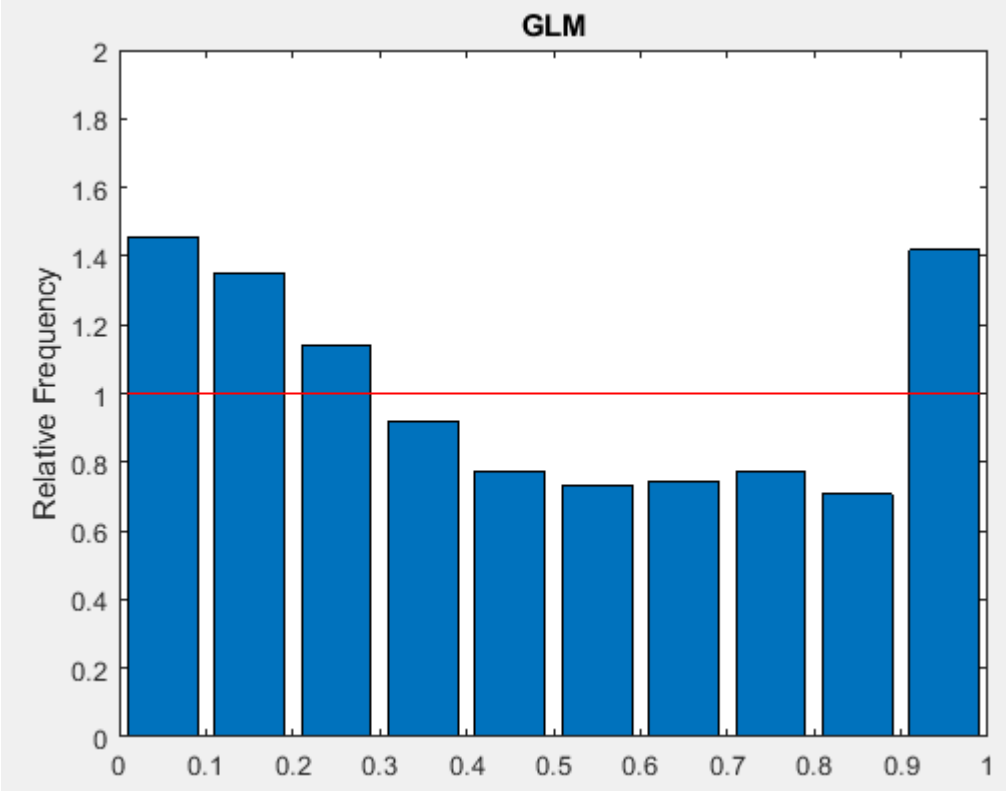

Figure 5.9: PIT histogram of GLM for the PhD articles data set.

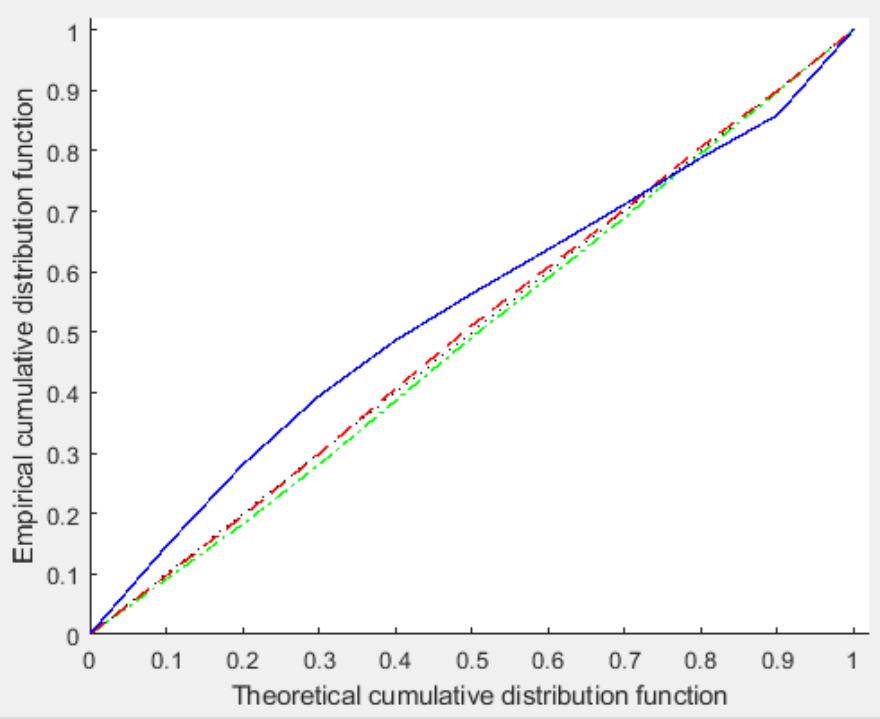

Figure 5.10: The P-P plots of PITs for SP-GLM-I (dashed-dotted green), SP-GLM-CL (dashed red) and GLM (solid blue) with the PhD articles data set.

5.3 CD4 data

In this section, we show the simplicity and effectiveness of the SP-GLM-CL method when the response distribution and the relationship between the response variable and the covariates are unknown and not suggested by intuition.

We apply the SP-GLM-CL method to find a relationship between the average of CD4 count and child's age. This data set was described in Wade and Ades (1994) and was provided in the R package `gamlss.data` (Rigby & Stasinopoulos, 2005). This data set contains n = 609 observations, with no missing data. The response variable is `cd4` – the CD4 counts of uninfected children born to HIV-1 infected mother, and the predictor variable is `age` – the child's age, in years.

Figure 5.11 (a) shows the histogram of CD4 counts. `cd4` is a discrete variable (count). `age` is a continuous variable and its histogram is shown in Figure 5.11 (b). The scatter plot of CD4 counts and child's age shown in Figure 5.12 suggests the possibility of a non-linear relationship between the two variables and the variation in `cd4` decreases with increasing `age`.

When fitting the model, SP-GLM-CL is fitted without any specification of the response distribution and link function. For GLM, we fit two models using (1) the Poisson distribution with log link (Poi-log) and (2) the normal distribution with identity link (N-id). We first fit the data using the Poisson regression model since `cd4` is a count response. For the second regression model, we assume that `cd4` has a continuous distribution since it is large enough to be treated as a continuous variable and the observed response values range from 0 to $2,327$. We build the model with the normal distribution and identity link as `cd4` contains zero and it is a convenient way to fit a model. We have tried to fit the normal regression model with log and square root link functions to the data, however, these models did not converge.

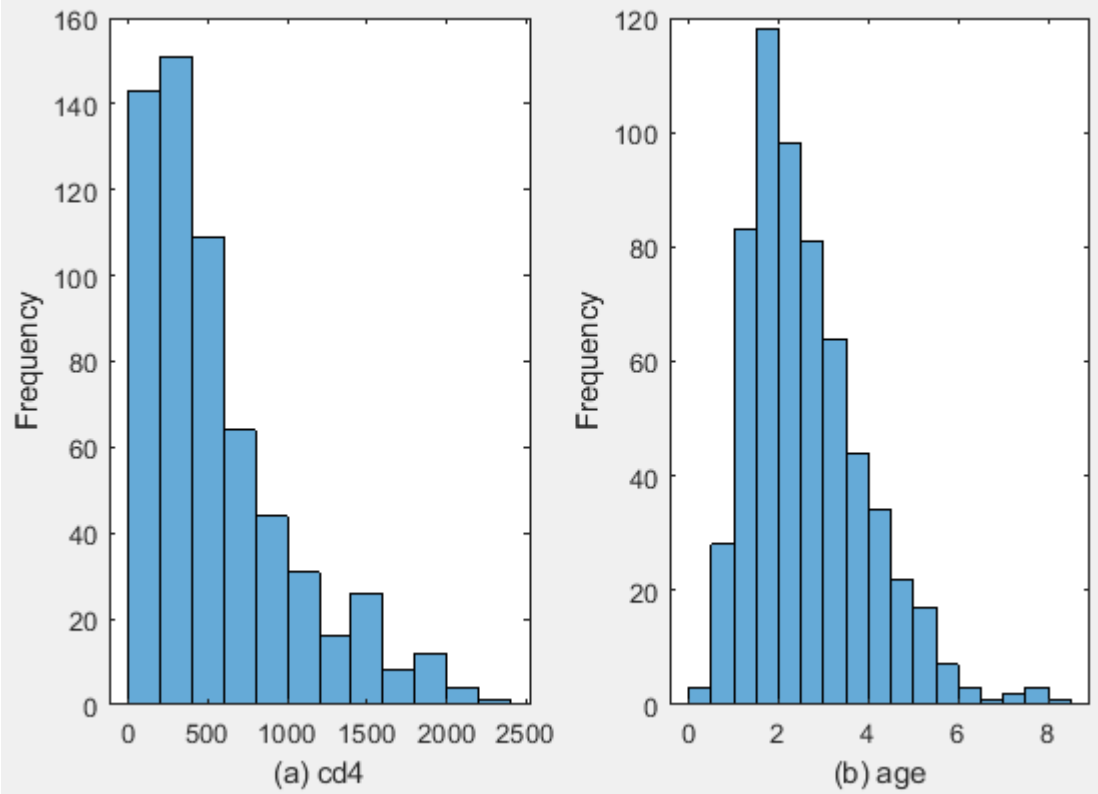

Figure 5.11: Histogram of cd4 and age.

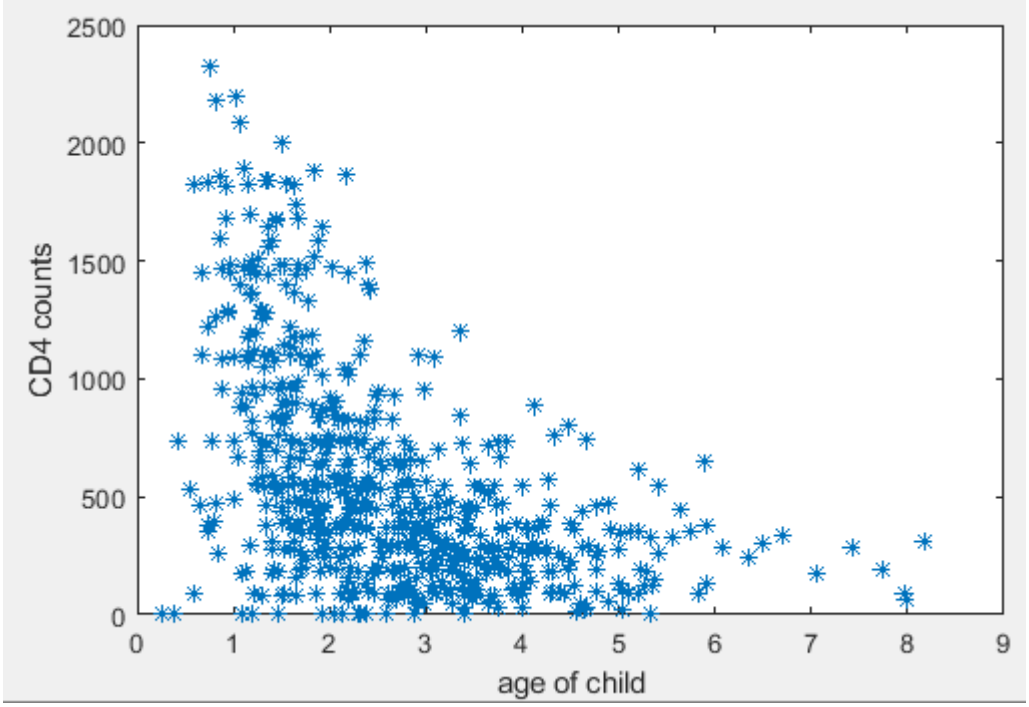

Figure 5.12: Scatter plot of CD4 counts and child's age.

We evaluate the estimated regression coefficients and the coefficient standard errors for each model. The fit of the models are assessed using the mean squared error (MSE) of the predictor that is obtained from $\sum_{i=1}^{n} [\texttt{cd4}_i - \hat{\mu}_i]^2 / n$ where $\hat{\mu}_i$ is the fitted value of the response $\texttt{cd4}_i$. We also compare the MSE of the predictor using the leave-one-out cross- validation method (MSE-CV) for the SP-GLM-CL, Poi-log and N-id models. The P-P plot of PIT and the PIT histogram are used for diagnostic checking for the model fit and the underlying distribution in each model.

The fitted model results are provided in Table 5.6. However, the comparison of the coefficient estimates and their standard errors in each model may not make sense since the response distribution and link function of SP-GLM-CL are not specified. The implicit canonical link function used in the SP-GLM-CL model can be observed from the plot of the linear predictor and the fitted mean. Figure 5.13 suggests that the SP-GLM-CL model uses the log link in model fitting for this data set. Even though the log link is used in both the SP-GLM-CL and Poi-log models, it is clearly seen from the coefficient estimates that the distribution used in the SP-GLM-CL model is not the Poisson distribution.

Table 5.6: The final model results of the CD4 data set.

	SP-GLM-CL		Poi-log		N-id	
Parameter	Estimate	ASE	Estimate	ASE	Estimate	ASE
intercept	-	-	7.2479	0.0039	1,007.70	36.78
age	-0.0016	0.0001	-0.3956	0.0017	-171.72	12.56

The fitted line shown in Figure 5.14 shows a better fit to the data for SP-GLM-CL (dashed red line) and Poi-log (dashed-dotted green line) than for N-id (solid black line). The poor fit of the N-id model may due to the constant variance assumption which does not agree with the data. This is supported by the MSE and MSE-CV values as shown in Table 5.7, where

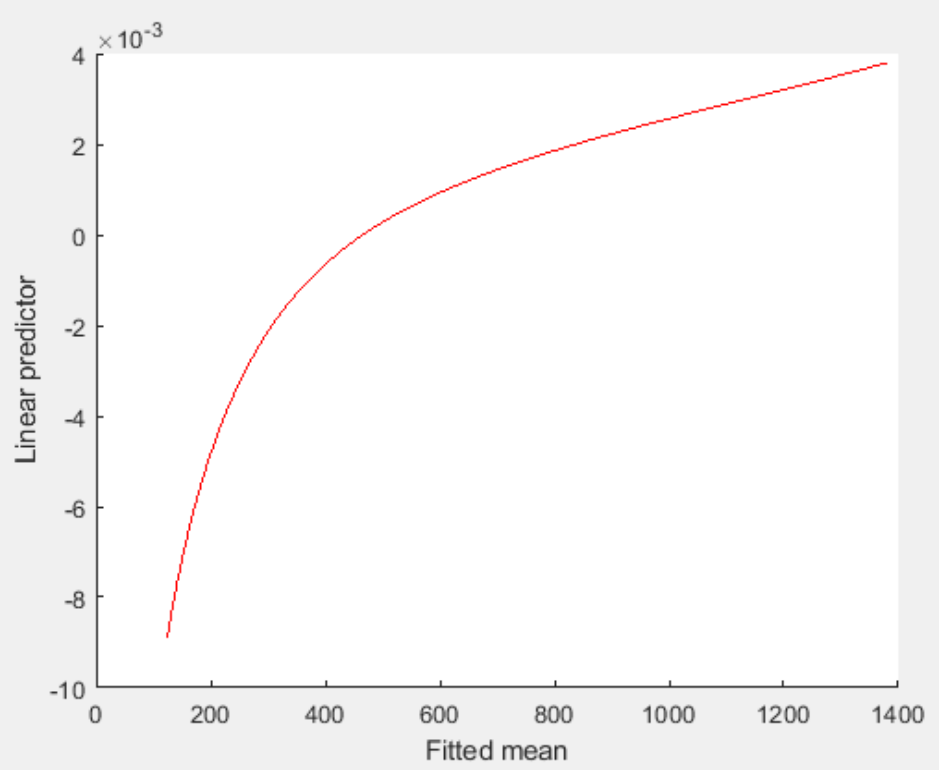

Figure 5.13: The fitted mean curve of SP-GLM-CL model for the CD4 data set.

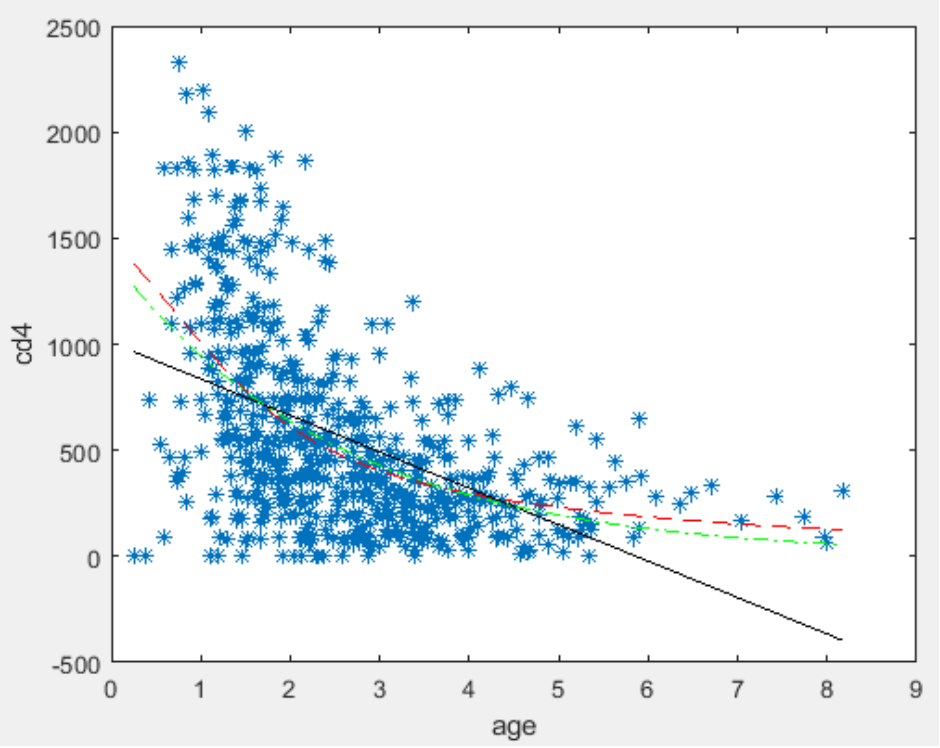

Figure 5.14: The scatter plot of `cd4` and `age` and the fitted curves for SP-GLM-CL (dashed red), Poi-log (dashed-dotted green) and N-id (solid black).

Table 5.7: The mean squared error (MSE) and the MSE using the leave-one-out cross-validation method (MSE-CV) for the model fitting of the cd4 data set.

	SP-GLM-CL	Poi-log	N-id
MSE	149,440	151,476	163,257
MSE-CV	150,835	152,739	164,459

both SP-GLM-CL and Poi-log models have lower values than the N-id model. However, the SP-GLM-CL method has the lowest MSE and MSE-CV values. Thus the SP-GLM-CL model provides better fit to the data than the Poi-log and N-id models.

When we assess the model fit and the underlying distribution in model fitting, only the SP-GLM-CL model has a good model fit and a proper response distribution. This is indicated by the histogram of the PITs shown in Figures 5.15 - 5.17. While the PIT histogram of the SP-GLM-CL model (Figure 5.15) is close to uniform, the plot of the PITs for the Poi-log model (Figure 5.16) shows an obvious U-shaped and a curve pattern has appeared in the plot for the N-id model (Figure 5.17). Moreover, the P-P plots of PITs in Figure 5.18 show that only the plot for SP-GLM-CL (dashed red line) is close to the 45° line (dotted black line), while the Poi-log (dashed-dotted green line) and N-id (solid blue line) plots are bent away from the comparison line.

The example in this section has demonstrated the usefulness and effectiveness of the proposed canonical link model fitting. The method is significantly simpler and easier to use for model fitting than the GLM method, where it is more cumbersome to find a suitable combination of the response distribution and the link function.

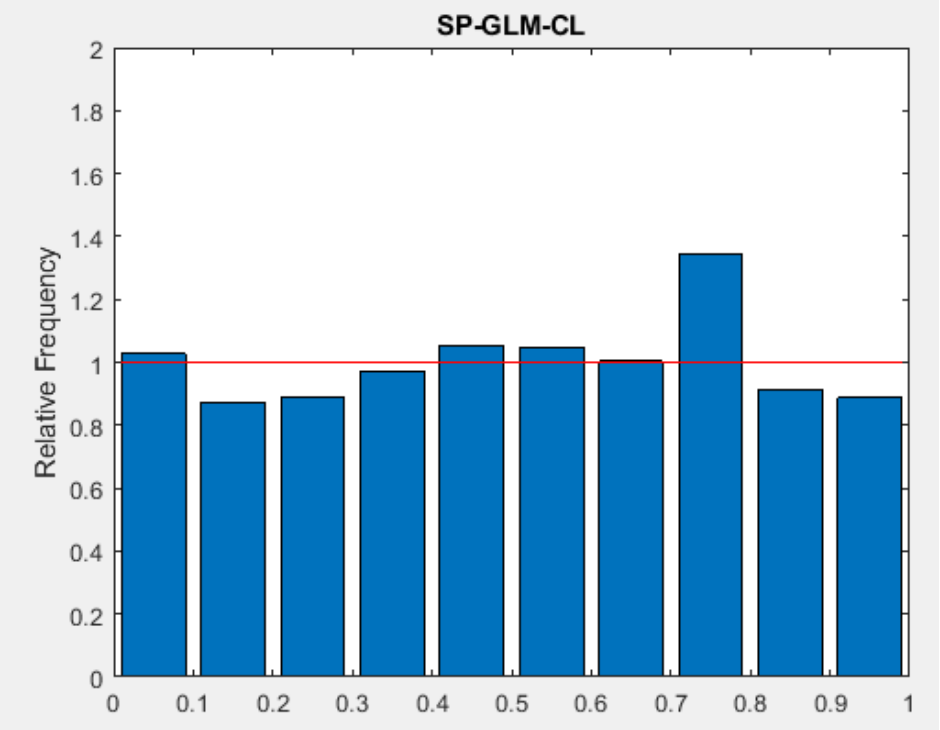

Figure 5.15: PIT histogram of SP-GLM-CL model for the CD4 data set.

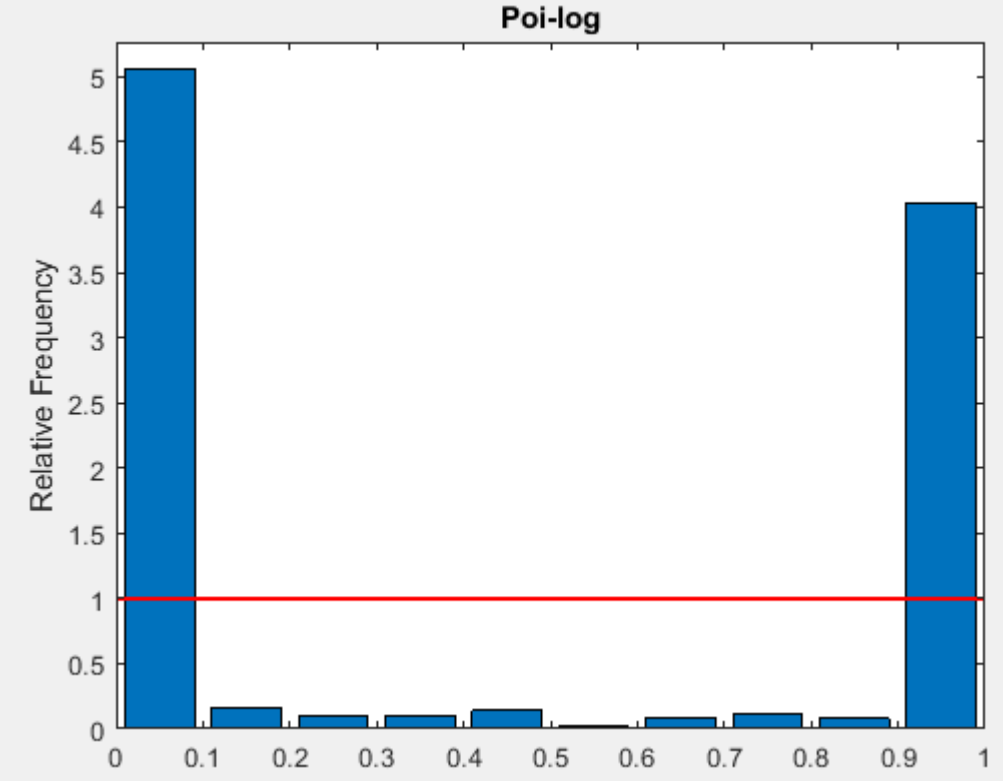

Figure 5.16: PIT histogram of Poi-log model for the CD4 data set.

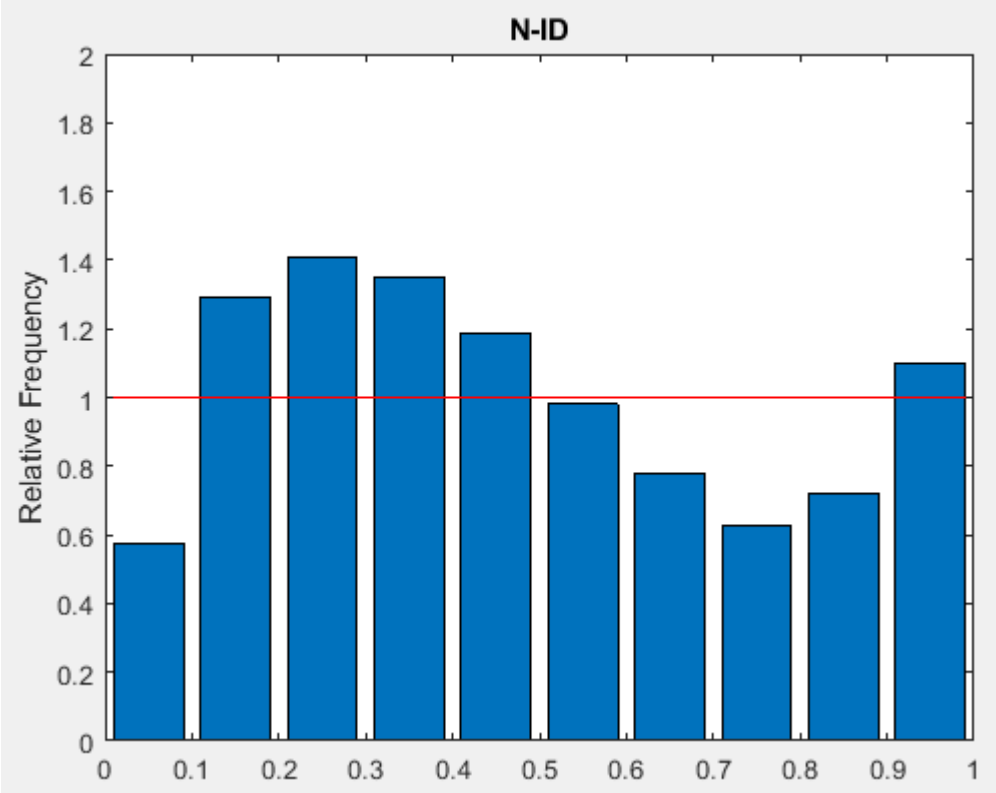

Figure 5.17: PIT histogram of N-id model for the CD4 data set.

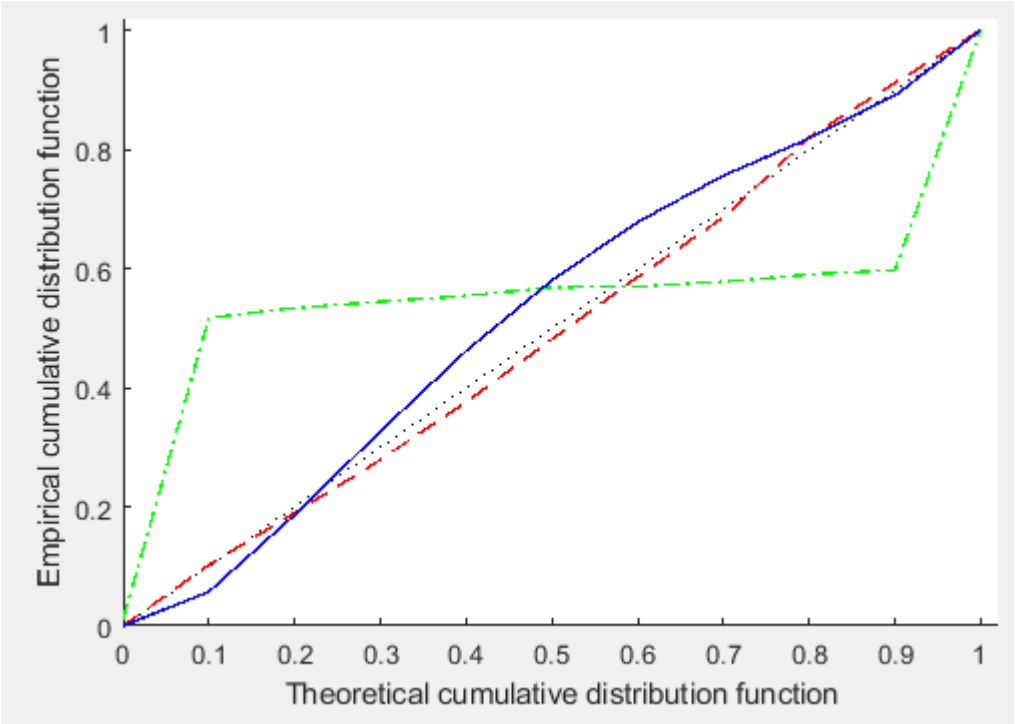

Figure 5.18: The P-P plots of PITs for SP-GLM-CL (dashed red), Poi-log (dashed-dotted green) and N-id (solid blue) with the CD4 data set.

5.4 Summary of applications to real data sets

These examples demonstrate increased effectiveness of the SP-GLM-I and SP-GLM-CL methods in practice compared with the GLM method. Both methods can handle very large data sets with several types of predictor variables. Their models usually show a better fit than GLM when the response distribution is beyond the set of distributions in the GLM framework. For SP-GLM, the response distribution is automatically selected by the data. In addition, when the response distribution and the link function are in doubt, the SP-GLM-CL method is very helpful to the modeller since neither the response distribution nor the link function are required to be specified and the SP-GLM-CL method usually shows better results than GLM in this situation.

6

Conclusions and Future Work

In this thesis, we have developed two novel methods to fit the SP-GLM with an unspecified reference density. The first method (SP-GLM-I) is developed for general cases in the SP-GLM, while the second one (SP-GLM-CL) is developed specifically for canonical link functions. This chapter provides summary of our proposed methods and possible future research directions.

6.1 Conclusions

The SP-GLM in this thesis refers to the one introduced by Rathouz and Gao (2009), that contains a nonparametric component in the response distribution. The SP-GLM provides greater flexibility for regression analysis than the traditional GLM, particularly when the response distribution is in doubt. To be clear, the SP-GLM offers great convenience of model fitting by eliminating the need to explicitly specify the response distribution. Such simplification provides great benefits to users/analysts, who may be unfamiliar with or inexperienced in model building.

However, the computational algorithm for the SP-GLM is more challenging than GLM. One reason for this is that the SP-GLM incorporates the unknown response density which is an infinite dimensional parameter. Moreover, the constraints on the response density make the model fitting algorithm more complex. The existing SP-GLM method (Huang, 2014) that relies on the built-in optimization function cannot handle large data sets and the model fitting algorithm is slow. Thus we have developed a novel model fitting method for the SP-GLM that is effective and reliable to use in practice. When using the first method (SP-GLM-I) to fit the SP-GLM, we iteratively estimate the regression coefficients together with the unknown reference density by maximizing the constrained log-likelihood function using the MI - Scoring algorithm.

For GLM, the canonical link function of the response variable is generally used in model fitting since the response mean is mostly guaranteed to be within the response's range. For the second method, we develop a novel fitting method for a special case of the SP-GLM when the canonical link is applied. Again, the regression coefficients are simultaneously estimated with the reference density through an iterative method. The Newton - MI algorithm is used to obtain these parameter estimates that maximize the constrained log-likelihood function.

The model fitting algorithm is simplified, and in general this method (SP-GLM-CL) can converge faster than the first method (SP-GLM-I). In terms of model building, this canonical link method (SP-GLM-CL) has demonstrated the ability to reduce the complexity of model building in GLM because it allows the model builder to bypass the cumbersome steps of working out suitable combinations of the response distribution and the link function, which would be required to ensure good model fitting in conventional GLM. In addition, the GLM method may fail to converge with some combinations of the specified response distribution and the link function.

In both methods, the reference density is approximated using piecewise constant functions. The constraints on the reference density are imposed via a Lagrange multipliers approach and the MI algorithm. The MI algorithm is used to estimate the reference density as it can impose the non-negativity constraints. This method also reduces the computational cost since the inverse of the information matrix is not required when estimating the reference density.

The consistency and asymptotic normality of the constrained maximum likelihood estimators for both methods are provided for a fixed number of bins. The asymptotic variances with active constraints are derived and the accuracy of the asymptotic standard errors of the regression coefficients have been verified through the simulation studies.

The simulation studies also show that both of the proposed methods provide accurate coefficient estimates and inferences. They converge fast and can deal with large data sets. They can work well with various types of response variables i.e. continuous, count and binary responses. Our model fitting methods are competitive with the existing SP-GLM method.

SP-GLM models provide comparable results to the standard GLM regression model if the response variable follows a distribution from the exponential family. In addition, they can provide effective results for fitting the model with a response variable that may have a

distribution outside the GLM framework, where standard GLMs may not be able to provide accurate parameter inferences. This is mainly due to the incorrect standard errors provided in GLM.

By applying our methods to real data, we were able to show that both of the proposed methods have good model fitting properties. They can handle the offset in the model. These methods can work well with various types of predictors in very large data sets. These applications show that an appropriate response distribution is automatically chosen in both models. In addition, the proposed methods can provide a better model fit than the GLM method when the distribution of the response is questionable.

Overall, both of our proposed model fitting methods have demonstrated their effectiveness when applied to simulated and real data sets. Both methods can be employed to enhance the use of the SP-GLM in practice, in particular through their ability to handle much larger data sets with significantly shorter computational time.

6.2 Future Work

A possible topic for future research is the application of the proposed SP-GLM-I and SP-GLM-CL model fitting methods on a random effects model such as the generalized linear mixed models (GLMMs) (McCulloch, 2003). GLMMs are extensions of GLMs where the random effects are included in the linear predictor besides the fixed effects. GLMMs provide an extension to the use of GLMs to allow for different distributions in the response variable. GLMMs can also handle correlated data such as the longitudinal data and grouped data. An example of the longitudinal data is the claims on one-year health insurance policies for each policyholder over some consecutive years. For a given policyholder, claims in consecutive years are correlated. Thus, a possibility exists to extend the SP-GLM-I and SP-GLM-CL

model fitting methods to include the random effects in the linear predictor, which would further broaden the types of data that the proposed methods can handle.

Another future direction is on improving estimation of the canonical parameter (θ_i). In the estimation of θ_i in Section 3.3.2, we reset the estimated response mean μ_i to stay within the possible range of $b'(\theta_i; \boldsymbol{p})$. In practice, this ad-hoc method can work well since most link function of interest satisfying the constraints on μ_i. However, we wish to explore in future how to impose the constraints on μ_i using constrained optimization methods.

Another possible direction for future work is the development of automatic selection of the optimal number of bins in the model. In the current model fitting methods, we use piecewise constant functions to approximate the reference density and the user is required to choose the number of bins. However, we wish to develop a procedure to automatically select the optimal number of bins.

We also wish to study in future extensions of the piecewise constant approximation of reference density approximation to other approximation methods.

All the computations for this thesis are made based on MATLAB. However, currently we are developing the `glmsp` R package. This new R package includes both model fitting methods proposed in this thesis, and it provides estimation results on regression coefficients and their standard errors, diagnostic check plots and likelihood ratio tests.

A

Appendix

This appendix contains the MATLAB code for the proposed methods presented in Chapter 3 (Appendix A.1) and Chapter 4 (Appendix A.2).

A.1 MATLAB code for SP-GLM-I

```
function [beta,den,other,SEb]=spglmI(y,X,distF,binC,maxI,link,ofs,b0)

% SPGLM_I: Semiparametric GLM fitting where regression coefficients
```

```
%          and reference density are estimated using the maximum
%          likelihood(ML).
% Outputs:
%    beta = vector of final regression coefficients estimiates
%    den = estimate of the discretized reference density
%    other = collection of iterations numbers and corresponding
%            log-likelihoods
%    SEb = estimate standard errors of regression coefficients
% Inputs:
%    y = a vertical n-vector
%    X = covariates matrix (n by q matrix) with intercept
%    distF = 0 if there is 1 observation in each bin;
%    distF = 1 if there is more than 1 observation in each bin
%    binC = number of observations in each bin
%    maxI = maximum number of iterations; default 500
%    link = link function
%    ofs = offset; default 0
%    b0 = initial value for beta; default 0
```

```
%% ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
y = y(:);
[n, q] = size(X);  % n = No. of rows; q = No. of columns
EY = mean(y) ;     % mean of y
% Initial beta
if b0==0
    % first value of initial beta
    if  strcmp(link,'identity')
        b1 = EY;
    elseif  strcmp(link,'log')
        b1 = log(EY);
    elseif  strcmp(link,'inverse')
        b1 = 1./EY;
```

```
    elseif  strcmp(link,'logit')
        b1 = log(EY/(1-EY))  ;
    elseif  strcmp(link,'sqrt')
        b1 = sqrt(EY) ;
    elseif  strcmp(link,'neginv')
        b1 = -1./EY;
    elseif  strcmp(link,'invsq')
        b1 = EY.^(-2);
    end
    beta0 = [b1; zeros(q-1, 1)];
else
    beta0 = b0;
end
```

```
%% Equal bin count
binid = 1:binC:n;
nbins = numel(binid);      % No. of bins
sy = sort(y);
binedg = (sy(binid))';     % vector of bin edges
cvg = zeros(maxI, 3);

i = 0;
while i < nbins
    i = i+1;
    ntie = sum(binedg(i)==binedg);
    if ntie > 1
        binedg = [binedg(1:i), binedg((i+ntie):nbins)];
        nbins = length(binedg);
    end
end
if binedg(end)==max(y) && nbins>1
    binedg(nbins+1) = max(y)+(binedg(nbins)-binedg(nbins-1));
```

```
    dbstop if error
else
    binedg(nbins+1) = max(y)+1e-5;
end
binwv = binedg(2:end)-binedg(1:end-1); % vector of bin width

% bin counts (No. of obs. in each bin)
[binCounts, index] = histc(y, binedg);
binCounts = binCounts(1:end-1);

% mid-point of bin
if distF==1      % if more than 1 obs. in each bin
    binM = binedg(1:nbins)+binwv/2;
elseif distF==0  % if 1 obs. in each bin
    binM = binedg(1:nbins);
end
binMm = repmat(binM,n,1);
maxYu = max(binM) ;  % max. value of mid-point of bin
minYu = min(binM) ;  % min. value of mid-point of bin
```

```
%% MI-Scoring iterations
denO = binCounts/n;   % initial reference density
theO = zeros(n, 1);
muXB = muXBeta(X, betaO, link, ofs); % initial mu
% reset \mu_i to boundary value
muXB(muXB<minYu) = minYu;
muXB(muXB>maxYu) = maxYu;
% initial Theta, bTheta, bpTheta, bppTheta
[theO,bthe,eytbt,mu,sig2]=Thetai(binMm,theO,denO,nbins,muXB,n);
% initial log-likelihood
llik0 = sum(y.*theO-bthe) + sum(binCounts.*log(denO));
```

```
for iter = 1:maxI
    %% MI algorithm for updating reference density
    sig2m = repmat(sig2,1,nbins);
    ymum = repmat(y-mu,1,nbins).*(binMm-repmat(mu,1,nbins));
    numer = binCounts./denO -(sum(min(0,ymum).*eytbt./sig2m))' ;
    denom = n + (sum(max(0,ymum).*eytbt./sig2m))'  ;
    % update reference density
    denN = denO .* (numer+0.1) ./ (denom+0.1);
    denInc = denN-denO;
    % update Theta, bTheta, bpTheta, bppTheta
    [theN,bthe,~,mu,sig2]=Thetai(binMm,theO,denN,nbins,muXB,n);
    % update log-likelihood
    llik1 = sum(y.*theN-bthe) + sum(binCounts.*log(denN));

    % line search
    ome = 0.6;
    while llik1 <= llik0
        denN = denO+ome*denInc;
        [theN,bthe,~,mu,sig2]=Thetai(binMm,theN,denN,nbins,muXB,n);
        llik1 = sum(y.*theN-bthe) + sum(binCounts.*log(denN));
        if ome >= 1e-4
            ome = ome*0.6;
        elseif ome < 1e-4 && ome >= 1e-6
            ome = ome*0.6^2;
        elseif ome < 1e-6 && ome >= 1e-20
            ome = ome*0.6^4;
        else
            break;
        end
    end
```

```
%% Fisher scoring algorithm for updating beta
% derivative of inverse link
if  strcmp(link,'identity')
    dgeta = ones(n,1);
elseif  strcmp(link,'log')
    dgeta = muXB ;
elseif  strcmp(link,'inverse')
    dgeta = - muXB.^2 ;
elseif  strcmp(link,'logit')
    dgeta = muXB.*(1-muXB) ;
elseif  strcmp(link,'sqrt')
    dgeta = 2* sqrt(muXB) ;
elseif  strcmp(link,'neginv')
    dgeta = muXB.^2 ;
elseif  strcmp(link,'invsq')
    dgeta = -0.5* muXB.^3 ;
end

W = (dgeta.^2)./sig2 ;
Fisher = zeros(q,q);
for i=1:n
    Xi = X(i,:);
    Fisheri = Xi'*W(i)*Xi;
    Fisher = Fisher + Fisheri;
end
V = dgeta./sig2 ;
Score = X' *((y - mu).*V);
binc = Fisher\Score;
% update beta
betaN = beta0 + binc;
% update mu
muXB = muXBeta(X, betaN, link, ofs);
```

```
muXB(muXB<minYu) = minYu;
muXB(muXB>maxYu) = maxYu;
% update Theta, bTheta, bpTheta, bppTheta
[theN,bthe,eytbt,mu,sig2]=Thetai(binMm,theN,denN,nbins,muXB,n);
% update log-likelihood
llik2 = sum(y.*theN-bthe)+sum(binCounts.*log(denN));

% line search
ome = 0.6;
while llik2 <= llik1
    betaN = beta0 + ome*binc;
    muXB = muXBeta(X, betaN, link, ofs);
    muXB(muXB<minYu) = minYu;
    muXB(muXB>maxYu) = maxYu;
    [theN,bthe,eytbt,mu,sig2]=Thetai(binMm,theN,denN,nbins,muXB,n);
    llik2 = sum(y.*theN-bthe) + sum(binCounts.*log(denN));
    if ome >= 1e-4
        ome = ome*0.6;
    elseif ome < 1e-4 && ome >= 1e-6
        ome = ome*0.6^2;
    elseif ome<1e-6 && ome>1e-20
        ome = ome*0.6^4;
    else
        break;
    end
end

%%
cvg(iter, :) = [iter, llik2, sum(denN)];
if all(abs(betaN-beta0)<1e-6) && all((abs(denN-den0))<1e-6)
    cvg = cvg(1:iter, :);
    break
```

```
    else
        beta0 = betaN;
        den0 = denN ;
        the0 = theN;
        llik0 = llik2;
    end
end
%%
beta = betaN ;
den = [binM', binwv', denN, binCounts, (1:nbins)'];
other.cvg = cvg;
other.fit = mu;
other.var = sig2;
phat = ones(n, nbins);
for i=1:n
    phat(i,:) = denN.*exp(binM'.*theN(i)-bthe(i));
end
```

```
%% Var-Cov matrix
% Fisher Information matrix
% derivative of inverse link
if  strcmp(link,'identity')
    dgeta = ones(n,1);
elseif  strcmp(link,'log')
    dgeta = muXB ;
elseif  strcmp(link,'inverse')
    dgeta = - muXB.^2 ;
elseif  strcmp(link,'logit')
    dgeta = muXB.*(1-muXB) ;
elseif  strcmp(link,'sqrt')
    dgeta = 2* sqrt(muXB) ;
elseif  strcmp(link,'neginv')
```

```
    dgeta = muXB.^2 ;
elseif  strcmp(link,'invsq')
    dgeta = -0.5* muXB.^3 ;
end
W = (dgeta.^2)./sig2 ;
Fisher = zeros(q,q);
for i=1:n
    Xi = X(i,:);
    Fisheri = Xi'*W(i)*Xi;
    Fisher = Fisher + Fisheri;
end
I11 = Fisher; % information matrix(beta)
denNm = repmat(denN',n,1);
mum = repmat(mu,1,nbins);
A = (binMidm-mum).*eytbt.*denNm;
V = diag((dgeta)./sig2) ;
I12 = -X'*V* A;
I21 = I12';
np2 = binCounts./(denN.^2) ;
I22_1 = zeros(nbins,nbins);
for i=1:n
    I22_1i = A(i,:)'/sig2(i)*A(i,:);
    I22_1 = I22_1 + I22_1i;
end
I22 = I22_1;
for id = 1:nbins
    I22(id,id) = I22(id,id)+np2(id);
end

% Fisher information matrix
Info = [I11, I12; I21, I22];
```

```
%% active constraint (pu=0 and sum(pu)=1)
sig2m = repmat(sig2,1,nbins);
ymum1 = repmat(y-mu,1,nbins);
% derivative of log-likelihood wrt. p
dldp=-(sum(ymum1.*(binMidm-mum).*eytbt./sig2m))'+binCounts./denN;
activecon2 = zeros(nbins,1);
activecon2(denN<1e-5 & dldp<1e-2 )=1; % active constraint pu=0
W1 = eye(nbins) ; % identity matrix for constrained pu=0
% choose row W1 only if active constrained pu=0
W2 = W1(activecon2==1,:);
W3 = [ones(1,nbins) ; W2] ; % add active constraint sum(pu)=1
rW = size(W3,1);
W4 = [zeros(rW,q), W3]; % add 0 for beta in active constraint
u4 = orth(W4');
Idbeta = [eye(q),zeros(q,nbins)] ;
% u'*u = Identity matrix
u = [Idbeta' , u4] ; % add column of Identity matrix for beta

% Asymptotic covariance matrix
Finv = u* inv(u'* Info * u )*u';

VarB = Finv(1:q,1:q); % cov for beta
SEb = diag(sqrt(VarB)); %standard error for beta
Varp = Finv(q+1:q+nbins,q+1:q+nbins);
SEp = diag(sqrt(Varp));
end
```

```
%% ++++++++++++++++++++++++++++++++++++++++++++++++++++++++
function [theN,bthe,eytbt,mu,sig2]=Thetai(binMidm,theO,denO,nbins,muXB,n
    )
% update Theta using Newton algorithm
theObc = theO ;
```

```
for itertheta = 1:30
    denOm = repmat(denO',n,1);
    theOm = repmat(theO,1,nbins);
    iebm = (exp(binMidm).^theOm).*denOm;
    eb = sum(iebm,2); eb(eb<=eps) = eps; eb(eb==Inf) = realmax;
    bthe = log(eb);    % bTheta
    bthem = repmat(bthe,1,nbins);
    eytbtp = exp(binMidm.*theOm - bthem).*denOm;
    mu = sum(binMidm.*eytbtp,2);  % bpTheta
    sig2 = sum((binMidm-repmat(mu,1,nbins)).^2.*eytbtp,2);% bppTheta
    sig2(sig2<eps) = eps; sig2(sig2==Inf) = realmax;
    inct = (mu - muXB)./ sig2 ;

    % number of integer digits of increment Theta
    inte = floor(log10(floor(abs(inct)))) +1 ;
    inte(inte==-Inf) = 0 ;
    w = 10.^(-inte) ; % weight for increment theta
    w(eb==realmax) =  0;
    theN = theO - w.*inct ;
    theN(eb==realmax) = theObc(eb==realmax);
    if all(abs(theO - theN)<1e-6) && all((abs(mu - muXB))<1e-6)
        break
    else
        theObc = theO ;
        theO = theN ;
    end
end
theNm = repmat(theN,1,nbins);
iebm = (exp(binMidm).^theNm).*denOm;
eb = sum(iebm,2); eb(eb<=eps) = eps; eb(eb==Inf) = realmax;
bthe = log(eb);
bthem = repmat(bthe,1,nbins);
```

```
eytbt = exp(binMidm.*theNm - bthem);
eytbtp = eytbt.*denOm;
mu = sum(binMidm.*eytbtp,2);
mu(mu==-Inf) = realmin; mu(mu==Inf) = realmax;
sig2 = sum((binMidm-repmat(mu,1,nbins)).^2.*eytbtp,2);
sig2(sig2<eps) = 1e-15; sig2(sig2==Inf) = realmax;
end
```

```
%% ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
function muXB = muXBeta(X, beta0, link, offset)
% Update mu
if  strcmp(link,'identity')
    muXB = X*beta0  + offset ;
elseif  strcmp(link,'log')
    muXB = exp(X*beta0  + offset);
elseif  strcmp(link,'inverse')
    muXB = 1./(X*beta0 + offset);
elseif  strcmp(link,'logit')
    muXB = exp(X*beta0 + offset)./(1+exp(X*beta0+offset));
elseif  strcmp(link,'sqrt')
    muXB = (X*beta0 + offset).^2 ;
elseif  strcmp(link,'neginv')
    muXB = -1./(X*beta0 + offset);
elseif  strcmp(link,'invsq')
    muXB = (X*beta0 + offset).^(-0.5) ;
end
end
```

A.2 MATLAB code for SP-GLM-CL

```
function [beta,den,other,SEb]=spglmCL(y,X,distF,binC,maxI,ofs)

% SPGLM_CL: Semiparametric GLM fitting with an unspecified
%           canonical link, where regression coefficients
%           and reference density are estimated using the
%           maximum likelihood(ML).
%Outputs:
%   beta = vector of final regression coefficients estimiates
%   den = estimate of the discretized reference density
%   other = collection of iterations numbers and corresponding
%           log-likelihoods
%Inputs:
%    y = a vertical n-vector
%    X = covariates matrix (n by q matrix) without intercept
%    distF = 0 if there is 1 observation in each bin;
%    distF = 1 if there is more than 1 observation in each bin
%    binC = number of observations in each bin
%    maxI = maximum number of iterations; default 500
%    ofs = offset; default 0

%% ++++++++++++++++++++++++++++++++++++++++++++++++++++++++

y = y(:);
[n, q] = size(X);   % n = No. of rows; q = No. of columns
tX = X;
meanX =  mean(X);   % mean of each X column
X = X - repmat(mean(X), n, 1); % centering X
beta0 = zeros(q, 1);
```

```
%% Equal bin count
binid = 1:binC:n;
nbins = numel(binid);  % No. of bins
sy = sort(y);
binedg = (sy(binid))'; % bin edges
cvg = zeros(maxI, 3);

i = 0;
while i < nbins
    i = i+1;
    ntie = sum(binedg(i)==binedg);
    if ntie > 1
        binedg = [binedg(1:i), binedg((i+ntie):nbins)];
        nbins = length(binedg);
    end
end
if binedg(end)==max(y)
    binedg(nbins+1) = max(y)+(binedg(nbins)-binedg(nbins-1));
else
    binedg(nbins+1) = max(y)+1e-5;
end
binwv = binedg(2:end)-binedg(1:end-1); % bin width
% mid-point of bin
if distF       % if more than 1 obs. in each bin
    binM = binedg(1:nbins)+binwv/2;
else           % if 1 obs. in each bin
    binM = binedg(1:nbins);
end
binMm = repmat(binM,n,1);
% bin counts (No. of obs. in each bin)
[binCounts, ~] = histc(y, binedg);
binCounts = binCounts(1:end-1);
```

```
%% Newton-MI iterations
den0 = binCounts/n;     % initial reference density
the0 = X*beta0 + ofs; % initial Theta
the0m = repmat(the0,1,nbins);
den0m = repmat(den0',n,1);

iebm = exp(binMm.*the0m).*den0m;
eb = sum(iebm,2); eb(eb<=eps) = eps; eb(eb==Inf) = realmax;
llik0 = sum(y.*the0-log(eb)) + sum(binCounts.*log(den0));

for iter = 1:maxI
    %% Newton algorithm for updating beta
    mu = sum(binMm.*iebm,2)./eb;
    sig2 = sum((binMm-repmat(mu,1,nbins)).^2.*iebm,2)./eb;
    sig2(sig2<eps) = eps; sig2(sig2==Inf) = realmax;
    WX = repmat(sqrt(sig2),1,q).*X;
    Score = X'*(y-mu);
    Fisher = WX'*WX;
    binc = Fisher\Score;
    % update beta
    betaN = beta0 + binc;
    % update Theta
    theN = X*betaN  + ofs;
    theNm = repmat(theN,1,nbins);
    iebm = exp(binMm.*theNm).*den0m;
    eb = sum(iebm,2); eb(eb<=eps) = eps;
    % update log-likelihood
    llik1 = sum(y.*theN-log(eb)) + sum(binCounts.*log(den0));

    % line search
    ome = 0.6;
```

```
    while llik1 <= llik0
        betaN = beta0 + ome*binc;
        theN = X*betaN  + ofs;
        theNm = repmat(theN,1,nbins);
        iebm = exp(binMm.*theNm).*denOm;
        eb = sum(iebm,2); eb(eb<=eps) = eps;
        llik1 = sum(y.*theN-log(eb)) + sum(binCounts.*log(denO));
        if ome >= 1e-4
            ome = ome*0.6;
        elseif ome < 1e-4 && ome >= 1e-6
            ome = ome*0.6^2;
        elseif ome<1e-6 && ome>1e-20
            ome = ome*0.6^4;
        else
            break;
        end
    end
```

```
    %% MI algorithm for updating baseline function
    % update reference density
    denN = binCounts./(sum(exp(binMm.*theNm)./repmat(eb,1,nbins)))';
    denN(denN<eps)=eps;
    denInc = denN-denO;
    denNm = repmat(denN',n,1);
    iebm = exp(binMm.*theNm).*denNm;
    eb = sum(iebm,2); eb(eb<=eps) = eps; eb(eb==Inf) = realmax;
    % update log-likelihood
    llik2 = sum(y.*theN-log(eb)) + sum(binCounts.*log(denN));

    % line search
    ome = 0.6;
    while llik2 <= llik1
```

```
        denN = denO+ome*denInc;
        denN = denN/sum(denN);
        denNm = repmat(denN',n,1);
        iebm = exp(binMm.*theNm).*denNm;
        eb = sum(iebm,2); eb(eb<=eps) = eps;
        llik2 = sum(y.*theN-log(eb)) + sum(binCounts.*log(denN));
        if ome >= 1e-4
            ome = ome*0.6;
        elseif ome < 1e-4 && ome >= 1e-6
            ome = ome*0.6^2;
        elseif ome < 1e-6 && ome >= 1e-20
            ome = ome*0.6^4;
        else
            break;
        end
    end
    %%
    cvg(iter, :) = [iter, llik2, sum(denN)];
    if all(abs(betaN-betaO)<1e-5) && all((abs(denN-denO))<1e-5)
        cvg = cvg(1:iter, :);
        break
    else
        betaO = betaN;
        denO = denN;
        denOm = denNm;
        llik0 = llik2;
    end
end
den = [binM', binwv', denN, binCounts];
other.cvg = cvg;
theNm = repmat(theN,1,nbins);
denNm = repmat(denN',n,1);
```

```
iebm = exp(binMm.*theNm).*denNm;
eb = sum(iebm,2); eb(eb<=eps) = eps;
mu = sum(binMm.*iebm,2)./eb;
sig2 = sum((binMm-repmat(mu,1,nbins)).^2.*iebm,2)./eb;
other.fit = mu; other.var = sig2;
bthe = log(eb);
phat = ones(n, nbins);
for i=1:n
    phat(i,:) = denN.*exp(binM'.*theN(i)-bthe(i));
end
```

```
%% Var-Cov matrix
% Fisher Information matrix
ebm = repmat(eb,1,nbins);
eytbt = exp(binMm.*theNm)./ebm;
WX = repmat(sqrt(sig2),1,q).*X;
Fisher = WX'*WX;
I11 = Fisher; % information matrix(beta)
mum = repmat(mu,1,nbins);
A = (binMm-mum).*eytbt;
I12 = X'* A;
I21 = I12';
np2 = binCounts./(denN.^2) ;
I22_2 = -(eytbt'*eytbt);
I22 = I22_2;
for id = 1:nbins
    I22(id,id) = I22(id,id)+np2(id);
end

% Fisher information matrix
Info = [I11, I12; I21, I22];
```

```
%% active constraint (pu=0 and sum(pu)=1)
% derivative of log-likelihood wrt. p
dldp = -eytbt' + binCounts./denN;
activecon2 = zeros(nbins,1);
activecon2(denN<1e-5 & dldp<1e-2 )=1; % active constraint pu=0
W1 = eye(nbins) ; % identity matrix for constrained pu=0
% choose row W1 only if active constrained pu=0
W2 = W1(activecon2==1,:);
W3 = [ones(1,nbins) ; W2] ; % add active constraint sum(pu)=1
rW = size(W3,1);
W4 = [zeros(rW,q), W3]; % add 0 for beta in active constraint
u4 = orth(W4');
Idbeta = [eye(q),zeros(q,nbins)] ;
% u'*u = Identity matrix
u = [Idbeta' , u4] ; % add column of Identity matrix for beta

% Asymptotic covariance matrix
Finv = u* inv(u'* Info * u )*u';

VarB = Finv(1:q,1:q); % cov for beta
SEb = diag(sqrt(VarB)); %standard error for beta
Varp = Finv(q+1:q+nbins,q+1:q+nbins);
SEp = diag(sqrt(Varp));
end
```

References

Aeberhard, W. H., & Hannay, M. (2018, Jun. 21). *Efficient semi-parametric generalized linear models based on exponentially tilted splines [abstract].* Presented at: The 2nd International Conference on Econometrics and Statistics. Retrieved from http://cmstatistics.org/RegistrationsV2/EcoSta2018/viewSubmission.php?in=409&token=8s2s6rrsss52312342n2r1475poq31r8

Asmussen, S., & Glynn, P. W. (2007). *Stochastic simulation: algorithms and analysis* (Vol. 57). Springer Science & Business Media.

Barndorff-Nielsen, O., & Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 279–312.

Blough, D. K., Madden, C. W., & Hornbrook, M. C. (1999). Modeling risk using generalized linear models. *Journal of health economics*, *18*(2), 153–171. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167629698000320 doi: https://doi.org/10.1016/S0167-6296(98)00032-0

Breheny, P. (2013, January). *Exponential families.* University Lecture Notes. Retrieved 1 March 2018, from https://web.as.uky.edu/statistics/users/pbreheny/760/S13/notes/1-31.pdf

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms: 2. the new algorithm. *IMA Journal of Applied Mathematics*, *6*(3), 222–231.

Chen, J. (2010). *Semiparametric methods for the generalized linear model* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University.

Cizek, P., Härdle, W. K., & Weron, R. (2005). *Statistical tools for finance and insurance*.

Springer Science & Business Media.

Crowder, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory*, *2*(3), 305–330.

Cruz, M. G., Peters, G. W., & Shevchenko, P. V. (2015). *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk*. John Wiley & Sons.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, *65*(4), 1254–1261.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 631–650.

De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data* (Vol. 10). Cambridge University Press Cambridge.

Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*(4), 1231–1236. Retrieved from `http://www.jstor.org/stable/2532266`

Esscher, F. (1932). On the probability function in the collective theory of risk. *Scandinavian Actuarial Journal*, *1932*(3), 175-195. Retrieved from `http://dx.doi.org/10.1080/03461238.1932.10405883` doi: 10.1080/03461238.1932.10405883

Fahrmeir, L., & Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, *13*(3), 317–322.

Frees, E. W., Derrig, R. A., & Meyers, G. (2014). *Predictive modeling applications in actuarial science* (Vol. 1). Cambridge University Press.

Fung, T., & Huang, A. (2016). Semiparametric generalized linear models for time-series data. *ArXiv e-prints*.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological bulletin*, *118*(3), 392–404.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, *24*(109), 23–26.

Haberman, S., & Renshaw, A. E. (1996). Generalized linear models and actuarial science. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *45*(4), 407–436. Retrieved from http://www.jstor.org/stable/2988543

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman & Hall/CRC.

Huang, A. (2014). Joint estimation of the mean and error distribution in generalized linear models. *Journal of the American Statistical Association*, *109*(505), 186–196.

Jackman, S. (2017). pscl: Classes and methods for R developed in the political science computational laboratory [Computer software manual]. Sydney, New South Wales, Australia. Retrieved from https://github.com/atahk/pscl/ (R package version 1.5.2)

Kafková, S., Křivánková, L., et al. (2014). Generalized linear models in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *62*(2), 383–388.

Kuhn, H., & Tucker, A. (1951). Nonlinear programming. In *Proceedings of the second berkeley symposium on mathematical statistics and probability* (pp. 481–492). Berkeley, California: University of California Press. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200500249

Lindsey, J. K. (1997). *Applying generalized linear models*. Springer.

Long, J. S. (1990). The origins of sex differences in science. *Social Forces*, *68*(4), 1297–1316. Retrieved from http://www.jstor.org/stable/2579146

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage Publications.

Luenberger, D. G., & Ye, Y. (1984). *Linear and nonlinear programming* (Vol. 2). Springer.

Ma, J. (2010, Feb). Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction. *IEEE Transactions on Nuclear Science*, *57*(1), 181-192. doi: 10.1109/TNS.2009.2034462

Ma, J., Couturier, D.-L., Heritier, S., & Marschner, I. (2017). *Penalized likelihood estimation for semiparametric proportional hazard models with interval censored data.* Unpublished manuscript, Macquarie University, Department of Statistics.

Ma, J., Heritier, S., & Lô, S. N. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Computational Statistics & Data Analysis*, *74*, 142–156.

Manning, W. G., & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of health economics*, *20*(4), 461–494. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0167629601000868` doi: https://doi.org/10.1016/S0167-6296(01)00086-8

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. London: Chapman & Hall.

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

McCulloch, C. E. (2003). Generalized linear mixed models. *NSF-CBMS Regional Conference Series in Probability and Statistics*, *7*, i–84. Retrieved from `http://www.jstor.org/stable/4153190`

Mihaylova, B., Briggs, A., O'hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health economics*, *20*(8), 897–916.

Moore, T. J., Sadler, B. M., & Kozick, R. J. (2008, March). Maximum-likelihood estimation, the cramér - rao bound, and the method of scoring with parameter constraints. *IEEE Transactions on Signal Processing*, *56*(3), 895-908. doi: 10.1109/TSP.2007.907814

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models* (Vol. 21). Springer.

Owen, A. B. (2001). *Empirical likelihood*. Chapman & Hall/CRC.

Rathouz, P. J., & Gao, L. (2009). Generalized linear models with unspecified reference distribution. *Biostatistics*, *10*(2), 205–218.

Renshaw, A. E., & Haberman, S. (1986). Statistical analysis of life assurance lapses. *Journal of the Institute of Actuaries*, *113*(3), 459–497. doi: 10.1017/S0020268100042566

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, *54*, 507-554.

Rodriguez, G. (2007). *Generalized linear models.* University Lecture Notes. Retrieved 1

March 2018, from http://data.princeton.edu/wws509/notes/

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, *11*(4), 735–757.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, *24*(111), 647–656.

Sun, W., & Yuan, Y.-X. (2006). *Optimization theory and methods: nonlinear programming* (Vol. 1). Springer Science & Business Media.

Thisted, R. A. (1988). *Elements of statistical computing: Numerical computation*. London, UK, UK: Chapman & Hall, Ltd.

Wade, A., & Ades, A. (1994). Age-related reference ranges: significance tests for models and confidence intervals for centiles. *Statistics in medicine*, *13*(22), 2359–2367.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, *61*(3), 439–447.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. Retrieved from http://www.jstor.org/stable/1912526

Wurm, M. J., & Rathouz, P. J. (2018). Semiparametric Generalized Linear Models with the gldrm Package [in press]. *The R Journal*. Retrieved from https://journal.r-project.org/archive/2018/RJ-2018-027/index.html (Available online on 2018-05-21)