Statistical Models for Speech Perception Tests

Wenli Hu

Master of Research

Department of Statistics

Faculty of Science

Macquarie University

October 2014

CONTENTS

ABSTRACT

Speech perception tests are applied to evaluate speech intelligibility of hearing-impaired subjects. Subjects in a sound laboratory listen to sentences, which they repeat back. The proportion of correctly identified test items for each sentence is the response of interest and lies between zero and one, with high frequencies at zero and one, corresponding to sentences completely misunderstood or correctly identified. Historically, these data have been analyzed in a two-step procedure. In the first step, responses in each block of sentences are aggregated to a single number, the speech perception threshold (SRT), which is the signal-to-noise ratio (SNR) at which the proportion correct is 0.5. In the second stage, SRTs are analyzed using analysis of variance techniques. We instead propose a zero-and-one inflated Beta regression model for proportion correct as the response. Several advantages are associated with the new approach: (1) complete data is employed; (2) the special distributional properties of proportion correct are accommodated in our inflated Beta distribution; (3) more inferences can be generated than that of a single threshold point; (4) random effects of subjects are considered. The proposed model is successfully applied to two sample data sets from studies, conducted by Cochlear, the company which manufactures the cochlear implants.

DECLARATION

The work described in this thesis was carried out in the Department of Statistics, Faculty of Science, Macquarie University, Sydney, Australia.

This is to certify that the material presented in this thesis is original, and has not been submitted for a degree or diploma at a university or other institute of higher learning.

Wenli Hu

October 2014

ACKNOWLEDGEMENTS

My deepest gratitude is given to my advisor, Associate Professor Gillian Heller. She is an amazing mentor, and has given me so much guidance, belief and encouragement. I would no have been able to advance step by step academically and finish this thesis without her supports. It has been a great learning experience. I also want to express my gratitude towards my co-supervisor, Dr. Brett Swanson, for this wonderful project, and for sharing his expertise in the hearing field. Thanks are also given to the Cochlear company, for providing the data sets used in the study.

I am indebted to my dear family, especially my mom, Huanying Wang, my dad, Aiwu Hu, and my brother, Shangwen Hu, for their endless and selfless love and supports. I am also grateful to my husband, Dr. Min Li, for making every moment memorable. Special mention also goes to two of my lifetime friends, from our days at Penn State: Dr. Jingyuan Liu and Dr. Wei Zhong, who always care for me, and are also guiding me towards the academic path in statistics.

Thanks to every member at the Department of Statistics, Macquarie University and indeed I appreciate every moment here.

1. INTRODUCTION

1.1 Models For Bounded Response Variables

Linear regression models the mean of the response variable y as a linear predictor, that is,

$$E(y) = \mu = x^T\beta$$

and assumes normally distributed errors with a constant variance. For situations where the outcome variable is bounded to an interval, the application of simple linear regression is no longer appropriate because the normality assumption is clearly not valid and it may produce fitted values that are out of the range.

Examples of bounded responses include rates and proportions that are restricted to the unit interval, which are frequently encountered in practice. For instance, in speech perception tests, one common measure of subject performance is the word score, defined as the proportion of correct words identified in a sentence and of the form $y = r/N$, where r is the number of words recognized correctly and N the total number of words in the sentence. In cases when the outcome is of this form, and words are independent, classical binomial modeling is applicable and available in the generalized linear modeling (GLM) framework. However when words are not independent, binomial modeling is no longer applicable. The binomial model is also no longer appropriate when r and N are unknown, yet the response is restricted to (0,1) and alternative approaches are required in terms of modeling.

Appropriate methods for the analysis of bounded outcomes are of great interest but the literature on this topic is relatively scarce. In a comprehensive review, Kieschnick and McCullough (2003) discussed from past literature some alternatives that have been adopted in the modeling of bounded responses. Specifically, they compared seven approaches divided into two categories: parametric and quasi-parametric, depending on likelihood specifications. Based on this, they recommended either parametric regression established on the Beta distribution, or quasi-likelihood regression. They further pointed out that the validity of asymptotic properties of quasi-likelihood estimates requires fairly large sample sizes, and the parametric Beta regression should be preferred unless the data satisfies the requirement of sample size.

In sections 1.1.1 to 1.1.4 we introduce several approaches that are applicable to the modeling of bounded responses on the unit interval. For situations where the response is bounded to an interval other than the standard unit, it can be transformed to support on [0,1], that is, if $u \in [a, b]$ and a, b are known with $a < b$, we have $u' = \left(\frac{u-a}{b-a}\right) \in [0, 1]$. To encompass infrequent occurrences of extreme values of 0 and 1, slight modification is employed by taking $y = \left[u'(N-1) + 0.5\right]/N$, where $y \in (0, 1)$ and N is the sample size (Smithson and Verkuilen 2006). Methods developed for outcomes limited to the open unit interval apply accordingly. However, when extreme values of 0 and 1 occur frequently in an informative way, inflated models can be considered.

1.1.1 Logit Normal Regression

One approach to analyze bounded responses is to transform the outcome variable, mapping its values to the real line, such that the transformed outcome could be modeled as a linear combination of independent variables as in classical regression. A popular choice is the logit transformation:

$$w = \log\left(\frac{y}{1-y}\right)$$

which yields $-\infty < w < \infty$. Assuming $w \sim \mathrm{N}\left(\mu, \sigma^2\right)$ implies that y has the Logitnormal distribution (LN) with density function

$$f\left(y; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma y\left(1-y\right)} \exp\left(-\frac{1}{2}\left(\frac{\log\left(\frac{y}{1-y}\right) - \mu}{\sigma}\right)^2\right) ,$$

which we denote as $y \sim \mathrm{LN}\left(\mu, \sigma^2\right)$. A regression approach is then to model the mean of w:

$$E\left(w\right) = E\left[\log\left(\frac{y}{1-y}\right)\right] = x^T\beta \ .$$

Lesaffre, Rizopoulos and Tsonaka (2007) proposed an extension to the above approach where they concentrated on the discrete bounded response on [0,1], in lieu of the mixed continuous-discrete type. It is assumed that a latent variable, denoted by u, is on the open interval (0,1) and has a Logitnormal distribution, namely, $\log\left(\frac{u}{1-u}\right) \sim \mathrm{N}\left(\mu, \sigma^2\right)$. Specifically, two cases were considered.

In the first case, the bounded outcome is a binomial proportion, where the true probabilities assume the Logitnormal distribution on (0,1). This approach is named the "binomial-logit-normal (BLN) approach". That is, the discrete response is of the form $y = r/N$ where $Ny \sim \text{Bin}\,(N, u)$ and u is the true probability, with $u \sim \text{LN}\,(\mu, \sigma^2)$. Replacing μ with $x^T\beta$ specifies the regression structure in modeling and we now have

$$E\left[\log\left(\frac{u}{1-u}\right)\right] = x^T\beta + \sigma z \ ,$$

where $z \sim \text{N}\,(0,1)$ is a random effect. Such a model could be analyzed as a "generalized linear mixed-effects model".

In the second case, the bounded outcome on [0,1] is discrete but not a proportion. It can be viewed as a coarsened version of u which is modeled through a "coarsening (CO) approach". For example, if y is produced in a way such that $y = k/m$ when $\frac{k-0.5}{m} \leq u < \frac{k+0.5}{m}$ with $k = 0, 1, ..., m$, we can define a set of boundary values as

$$a_0 = 0 < a_1 = \frac{0.5}{m} < .... < a_m = \frac{m-0.5}{m} < a_{m+1} = 1 \ .$$

The response y is a discretized version of u and it is assumed that $a_{s(i)} \leq u_i < a_{s(i)+1}$ when y_i is observed (Lesaffre, Rizopoulos, and Tsonaka 2007). Therefore the likelihood is given by

$$L\left(y; \mu, \sigma^2\right) = \prod_{i=1}^{n} \int_{a_{s(i)}}^{a_{s(i)+1}} \frac{1}{\sqrt{2\pi}\sigma u_i\left(1-u_i\right)} \exp\left(-\frac{1}{2}\left(\frac{\log\left(\frac{u_i}{1-u_i}\right) - \mu_i}{\sigma}\right)^2\right) du \ .$$

Replacing μ with the linear predictor $x^T\beta$, the likelihood is

$$L\left(y; \beta, \sigma^2\right) = \prod_{i=1}^{n}\left(\Phi\left(\frac{\text{logit}\left(a_{s(i)+1}\right) - x_i^T\beta}{\sigma}\right) - \Phi\left(\frac{\text{logit}\left(a_{s(i)}\right) - x_i^T\beta}{\sigma}\right)\right) \ ,$$

where $\Phi\left(\cdot\right)$ is the distribution function of the standard normal distribution. Maximum likelihood estimates can be obtained.

Frederic and Lad (2008) commented on the differences between the Logitnormal distribution indexed with parameters μ and σ^2 and the Beta distribution indexed with α and β: "when σ^2 is small, the densities are unimodal and look similar to the Beta(α, β) densities when α and β both exceed 1. However, the Logitnormal densities always converge to 0 as $x \to 0^+$ and as $x \to 1^-$. In this way they differ markedly from the family of Beta(α, β) densities when α and β are both less than 1. These Beta densities have unbounded asymptotes at 0^+ and at 1^-. The Logitnormal family contains no members that resemble the Beta(α, β) densities when $\alpha \leq 1$ and $\beta \geq 1$, or vice versa". It can be readily seen that the logit transformation on data bounded to the standard unit interval would only work under limited conditions, when y assumes the Beta distribution. In addition, the true underlying probability distribution is rarely known. Moreover, when heteroscedasticity and asymmetry, two properties common in limited range variables, are prevalent and not addressed, the inferences based on the normality assumption are invalid and misleading. Furthermore, transformation poses difficulty in terms of interpretation and inferences on the original variable.

1.1.2 Logit Quantile Regression

An alternative to modeling bounded outcomes involves the application of logistic quantile regression to bounded continuous outcomes, which is valid with any population distribution (Bottai, Cai, and McKeown 2010). Let $Q_y(q)$ denote the q-th quantile of a random variable y restricted on (0,1). This approach assumes that with a suitable link function $g(\cdot)$, for every $Q_y(q)$, the coefficients of covariates are fixed and we have in the case of the logit link function

$$g\left(Q_y(q)\right) = \log\left(\frac{Q_y(q)}{1 - Q_y(q)}\right) = x^T\beta \ , \tag{1.1}$$

where x^T is a vector of covariates with coefficient vector β. Therefore, $Q_y(q)$ can be derived as

$$Q_y(q) = g^{-1}\left(x^T\beta\right) = \frac{\exp\left(x^T\beta\right)}{\exp\left(x^T\beta\right) + 1} \ , \tag{1.2}$$

which enables the comparison of distributions of the response for different covariate levels. The estimation of coefficients β is accomplished through expressing quantiles of the $g(\cdot)$ transformed response as a linear predictor, which is denoted by

$$Q_{g(y)}(q) = x^T\beta \ .$$

The applicability of this method, as well as the derivation of formulas (1.1) and (1.2), are based on the equality given by

$$Q_{g(y)}(q) = g\left(Q_y(q)\right) \ .$$

This states that with $g(\cdot)$, conditional quantiles of the transformed response are equivalent to the transformed conditional quantiles of the original response, a feature that is not applicable to the conditional mean (Mu and He 2007). The statement is valid because the relation $\Pr(y \leq y') = \Pr(g(y) \leq g(y'))$ exists for any random variable y

and any monotone, non-decreasing link function $g(\cdot)$. The disadvantage of this method is that the interpretations are made to quantiles and distributions in lieu of the original y.

1.1.3 Beta Regression

Beta Distribution

The Beta distribution is a continuous probability distribution of values supported on (0,1), indexed by two shape parameters. It is very flexible to work with in the sense that by adjusting the two parameters, variants of density shapes could be obtained such as U-shaped, uniform, reverse J-shaped, J-shaped, unimodal asymmetric and symmetric, as illustrated in Figure 1.1. The uniform distribution is a special case of the Beta distribution when the two shape parameters equal 1.

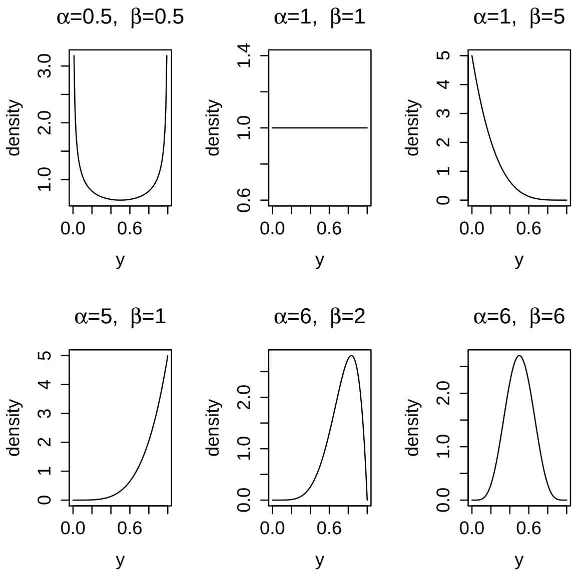


Fig. 1.1: Beta density plots for varying α and β.

The Beta distribution has probability density function

$$f(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \ , \tag{1.3}$$

where $\alpha > 0$, $\beta > 0$, $y \in (0,1)$ and $\Gamma(\cdot)$ is the gamma function. Under this parametrization, the mean and variance are given by

$$E(y) = \mu = \frac{\alpha}{\alpha+\beta} \ ,$$

$$Var\left(y\right)=\frac{\alpha\beta}{\left(\alpha+\beta\right)^{2}\left(\alpha+\beta+1\right)}=\frac{E\left(y\right)\left(1-E\left(y\right)\right)}{\alpha+\beta+1}\ .$$

It is noteworthy that the variance is a function of the mean.

If the probability distribution of a random variable y belongs to the exponential family, the probability density function (pdf) is of the form

$$f\left(y;\theta,\phi\right)=\exp\left(\frac{\eta\left(\theta\right)t\left(y\right)-B\left(\theta\right)}{a\left(\phi\right)}\right)h\left(y,\phi\right)\ ,$$

where $\eta\left(\theta\right)$ is the natural parameter of the exponential family, $t\left(y\right)$ is a sufficient statistic of θ and ϕ is the dispersion parameter. The Beta distribution, if one parameter is fixed, is a special case of the one-parameter exponential family, which can readily be seen from

$$f\left(y;\alpha,\beta\right)=\exp\left(\alpha\log\left(y\right)-\log\frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha+\beta\right)}\right)\frac{\left(1-y\right)^{\beta}}{y\left(1-y\right)\Gamma\left(\beta\right)} \tag{1.4}$$

$$=\exp\left(\beta\log\left(1-y\right)-\log\frac{\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)}\right)\frac{y^{\alpha}}{y\left(1-y\right)\Gamma\left(\alpha\right)}\ . \tag{1.5}$$

In (1.4), β is considered fixed and we have $\eta=\alpha$ and $t(y)=\log(y)$. In (1.5), α is considered fixed and accordingly, we have $\eta=\beta$ and $t(y)=\log(1-y)$. When both α and β are parameters, the Beta distribution belongs to the two-parameter exponential family with pdf

$$f\left(y;\alpha,\beta\right)=\exp\left(\alpha\log\left(y\right)+\beta\log\left(1-y\right)-\log\frac{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)}\right)\frac{1}{y\left(1-y\right)}\ ,$$

where we have $\eta=(\alpha,\beta)$, $t(y)=(\log(y),\log(1-y))$, $B(\theta)=\log\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, $h(y)=\frac{1}{y(1-y)}$.

It is well known that the application of GLM assumes the underlying probability distribution to be a member of the exponential family. Given that the Beta distribution is indeed one-parameter exponential when either α or β is fixed, or two-parameter exponential when both α and β are parameters, GLM on the basis of the Beta distribution is an applicable approach in which the response is restricted to the interval (0,1) and assumed to have the Beta distribution. This distributional property of the Beta distribution lays the foundation for Beta regression.

Beta Regression

Based on the assumption that the response assumes the Beta distribution, Beta regression is appropriate in situations when y is either of the form $y=r/N$ or not, through the mean response modeling as a linear combination of exogenous variables.

There exist in the literature different parametrizations indexing the Beta density. The standard approach modeled the shape parameters α, β in (1.3) as linear functions of covariates through a log link function:

$$\log(\alpha) = \eta = x^T\beta \ , \tag{1.6}$$

$$\log(\beta) = \zeta = z^T\gamma \ , \tag{1.7}$$

where x^T and z^T are p- and q-dimensional vectors of regressors and $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$, $\gamma = (\gamma_1, \gamma_2, ..., \gamma_q)^T$ are coefficients of those known covariates. Cribari-Neto and Vasconcellos (2002) used maximum likelihood to obtain estimates of α and β with bias corrections but without regression structures in terms of modeling; they further looked into regression analysis and described α and β as functions of a set of explanatory variables, as in formulas (1.6) and (1.7) (Vasconcellos and Cribari-Neto 2005). Paolino (2001) applied Monte Carlo simulation to show that the standard approach based on the Beta distribution generated more accurate and precise results than regular OLS method under the normality assumption. However a few criticisms were also raised, one of them being the difficulty in interpretation because inference was made on the shape parameters only. More often, the interest lies in the association between exogenous variables and the mean or variance. This leads to an alternative approach in terms of modeling the mean and dispersion in the Beta distribution.

In generalized linear modeling, it is always the mean μ that is modeled, which is of primary interest. In line with that idea, Ferrari and Cribari-Neto (2004) proposed a Beta regression model where the mean response is modeled through a parametrization in which the mean and dispersion characterize the Beta density. In particular, the Beta density is given by

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\,\Gamma((1-\mu)\,\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1} \ , \tag{1.8}$$

where $0 < \mu < 1$ and $\phi > 0$. Moreover, μ, ϕ and the original parameters α, β as in (1.3) are connected through the relations $\mu = \frac{\alpha}{\alpha+\beta}$, $\phi = \alpha + \beta$. The mean and variance are $E(y) = \mu$ and $Var(y) = \frac{\mu(1-\mu)}{1+\phi}$. ϕ is typically defined as a precision or dispersion parameter and specifically, when μ is fixed, larger ϕ results in smaller variance of y. Various probability density shapes with changing μ and ϕ are displayed in Figure 1.2.

For regression modeling of the mean, any link function that maps (0,1) to $(-\infty, \infty)$, i.e. any inverse sigmoidal function, could be appropriate. In practice one popular choice is the logit link function:

$$\log\frac{\mu}{1-\mu} = x^T\beta \quad \text{or} \quad \mu = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)} \ .$$

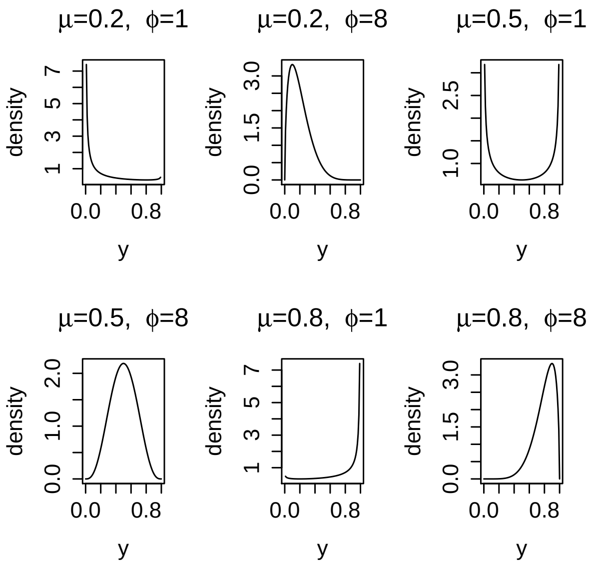


Fig. 1.2: Beta density plots for varying μ and ϕ.

The log-likelihood function is expressed as

$$\begin{aligned}\ell(\beta,\phi) &= \sum_{i=1}^{n} \log f(y_i;\mu_i,\phi) \\ &= \sum_{i=1}^{n} \left[\begin{array}{l} \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) \\ +(\mu_i\phi-1)\log y_i + ((1-\mu_i)\phi-1)\log(1-y_i) \end{array} \right].\end{aligned}$$

In this regression analysis, score functions and Fisher's information matrix were derived and iterative algorithms were applied in order to obtain maximum likelihood estimates. A complete set of inference tools was obtained including point and interval estimates, hypothesis testing and diagnostic measures. A toolbox is implemented in the R package "betareg" (Cribari-Neto and Zeileis 2010).

Classical Beta regression is further extended in various directions, one of which being the inclusion of a regression structure on the precision parameter ϕ along with more flexibility in regression structures. It takes into account the possibility that some covariates may affect the dispersion, causing additional heteroscedasticity than explained by the modeling of mean. The model equations are

$$\begin{aligned} g_1(\mu) &= \eta = x^T\beta \ , \\ g_2(\phi) &= \zeta = z^T\gamma \ , \end{aligned}$$

where x and z are p- and q-dimensional vectors of covariates and $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$, $\gamma = (\gamma_1, \gamma_2, ..., \gamma_q)^T$ are coefficients of those covariates. Note that the covariates enter the regression structures in a linear fashion (Smithson and Verkuilen 2006). Commonly used link functions are logit and log for the mean and precision parameter respectively.

Allowing parametric non-linear terms in the regression structures of both mean and precision further extends Beta regression (Simas, Barreto-Souza, and Rocha 2010). "Non-linear" indicates in the sense that the model is not linear in the coefficients. Some examples of parametric non-linear functions include the exponential function, logarithmic function or power function, to list a few. The mean and precision parameter could be expressed accordingly as

$$\begin{aligned} g_1(\mu) &= f_1\left(x^T;\beta\right) \ , \\ g_2(\phi) &= f_2\left(z^T;\gamma\right) \ , \end{aligned}$$

where $f(\cdot)$ denotes non-linear functions. Branscum, Johnson and Thurmond (2007) replaced the linear predictor of mean response with non-parametric spline functions in a semi-parametric Beta regression, where the analysis was in the context of Bayesian regression rather than maximum likelihood estimation. Specifically, the functional relation of μ was described as

$$g(\mu) = m(x) = \beta_0 + \beta_1 x + ...\beta_p x^p + \sum_{j=1}^{J} b_j \left((x-k_j)_+\right)^p \ ,$$

where the k_js are fixed knots and $(x)_+ = xI(x>0)$.

Another direction lies in the well established fact that the bias of maximum likelihood estimators (MLE) in small samples could be substantial and problematic, since the asymptotic properties of MLE are justified on the basis of large sample size. The bias of MLE is typically of the order $\mathrm{O}(n^{-1})$ and it is not of serious concern when sample size is relatively large, whereas in moderate or small samples, the bias needs to be taken into consideration. Ospina, Cribari-Neto and Vasconcellos (2006) adopted three approaches to address this issue: bias correction through the derivation of second-order bias; the preventive method of bias reduction proposed by Firth (1993); and a bootstrap resampling scheme, to obtain an estimation of bias. These measures all adjust MLEs for β and ϕ. Realizing the fact that the bias-free property of adjusted (β,ϕ) or (β,θ) in cases of modeling ϕ with covariates, to order $\mathrm{O}(n^{-1})$ does not necessarily apply to that of μ or (μ,ϕ), Simas, Barreto-Souza and Rocha (2010) further derived bias correction of MLEs of (μ,ϕ) when the precision parameter is also characterized by covariates.

A third direction involves Bayesian regression in which inference can be made from posterior samples. For instance, Branscum, Johnson and Thurmond (2007) applied Bayesian Beta regression with Markov chain Monte Carlo (MCMC) sampling to household expenditure data and generic distance between foot-and-mouth disease viruses. The likelihood function and the posterior distribution were defined as

$$L(\beta, \phi) = \prod_{i=1}^{n} \frac{\Gamma(\phi)}{\Gamma(\mu_i \phi)\,\Gamma((1-\mu_i)\,\phi)} y_i^{\mu_i \phi - 1} (1 - y_i)^{(1-\mu_i)\phi - 1} ,$$

$$p(\beta, \phi \,|y) = \frac{L(\beta, \phi)\, p(\beta, \phi)}{\int L(\beta, \phi)\, p(\beta, \phi)\, d\beta d\phi} ,$$

where informative prior distributions were constructed for β and ϕ respectively. Bayesian modeling is readily accessible through WinBUGS. It is noteworthy that the reliability of Bayesian regression relies on the specification of appropriate prior distributions, without which the inferences could be misleading.

Some other extensions include exploration on diagnostic measures, mixed Beta regression where discrepancies among groups with unobservable group memberships are accounted for (Verkuilen and Smithson 2012) and errors-in-variables Beta regression models where covariates are observed with error (Carrasco, Ferrari, and Arellano-Valle 2014), to name a few. Grün, Kosmidis and Zeileis (2012) incorporated bias correction/reduction of MLEs, Beta regression tree models and latent class Beta regression into Beta regression along with implementation of new functions into “betareg”. Other appealing research further extends the modeling capacity by enabling the incorporation of covariates that are parametric non-linear, random effects and non-parametric smoothing functions into the regression structures for both the mean and precision parameter. More importantly, it includes a rich number of other useful features, all implemented into the R package “gamlss” (“**G**eneralized **A**dditive **M**odels for **L**ocation, **S**cale and **S**hape”), which will be reviewed in the following section (Rigby and Stasinopoulos 2005).

More recently, Schmid et al (2013) proposed boosted Beta regression on the basis of an boosting algorithm called “gamboostLSS” motivated by “gamlss”. The introduction of the algorithm “gamboostLSS” was inspired by the finding that variable selection based on generalized AIC (GAIC) in “gamlss” has several shortcomings and is infeasible when more covariates exist than observations (Mayr et al 2012). Boosted Beta regression is aimed to address some limitations of classical Beta regression including variable selection issues, multicollinearity, non-linear predictor response relationships and covariate structures of the precision parameter. In particular, variable selection was achieved via a gradient boosting algorithm. Variable selection and parameter estimation were performed simultaneously, which was more stable than classical selection methods such as stepwise selection based on the values of some criteria.

1.1.4 Inflated Beta Regression

Response measures limited to the standard unit interval may encounter the extreme values of zero and/or one. Transformations of these into values in the open interval (0,1), followed by the application of aforementioned regression methods, results in a loss of information represented by those boundary observations. For instance, the rates of defect products occurring may naturally contain many zeros, which is suggestive of a good condition of manufacturing. If such information is ignored, the inferences generated would not be fully reflective. As a result, special considerations are required and further, methods based on inflated distributions are applicable.

The word "inflated" means that some values have more probability mass than explained by the proposed distribution. A well-known example is the zero-inflated Poisson distribution, where more zeros occur than accommodated by the Poisson distribution. The zero-inflated Poisson distribution has a long history, of which some early literature include Cohen (1963) and Johnson and Kotz (1969). The probability function is given by

$$f\left(y; p, \lambda\right) = \begin{cases} p + \left(1-p\right) e^{-\lambda} & y = 0 \\ \left(1-p\right) e^{-\lambda} \frac{\lambda_i^y}{y!} & y = 1, 2, ... \end{cases} ,$$

which is a mixture of a degenerate distribution at zero assigned a probability of p, and the Poisson distribution, assigned a probability of $(1-p)$. It is noteworthy that those earliest applications do not involve any regression structure of parameters. Lambert (1992) first applied zero-inflated Poisson regression in which parameters λ and p were modeled, to defects in manufacturing, where λ and p had functional forms as

$$\log\left(\frac{p}{1-p}\right) = x^T\beta \ ,$$
$$\log\left(\lambda\right) = z^T\gamma \ .$$

The likelihood function was further deduced, followed by EM algorithm implementation in order to obtain estimates. In addition to maximum likelihood estimation, Bayesian estimation, with specifications of prior distributions and the aid of MCMC embedded in WinBUGS, could be useful in applications as well (Ghosh, Mukhopadhyay, and Lu 2006). The idea of zero-inflated models is applicable to other probability distributions such as the negative binomial, gamma, or inverse Gaussian distributions, to name a few, under the condition that the assumed distribution fits the data scenario. In a similar fashion, one-inflated models could be analyzed for cases where the probability mass of 1 exceeds that permitted by the assumed distribution and is of meaningful insight.

We focus our attention on zero-and/or-one inflated Beta regression for observations on intervals [0,1),(0,1] and [0,1] separately, and in particular, related contributions made by Ospina & Ferrari (2010) and Rigby & Stasinopoulos (2005) . Ospina and Ferrari (2010) fully described the aforementioned three distributions, namely zero-inflated Beta distribution (BEZI), one-inflated Beta distribution (BEOI) and zero-and-one inflated Beta distribution (BEINF). With the same annotations used, if a random variable y assumes the zero-inflated Beta distribution, that is, $y \sim \mathrm{BEZI}\left(\alpha,\mu,\phi\right)$, the corresponding cumulative density function (cdf) of the mixed discrete-continuous distribution is described by

$$F_{BEZI}\left(y;\alpha,\mu,\phi\right)=\begin{cases}0 & y<0\\ \alpha & y=0\\ \alpha+\left(1-\alpha\right)\int_0^y \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}t^{\mu\phi-1}\left(1-t\right)^{(1-\mu)\phi-1}dt & y\in(0,1)\end{cases}$$

We write the mixed discrete-continuous probability function as

$$f_{BEZI}\left(y;\alpha,\mu,\phi\right)=\begin{cases}\alpha & y=0\\ \left(1-\alpha\right)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}\left(1-y\right)^{(1-\mu)\phi-1} & y\in(0,1)\end{cases}\tag{1.9}$$

where $0<\alpha<1$, $0<\mu<1$ and $\phi>0$. Figure 1.3 characterizes some shapes of the probability function, given a fixed α, for the zero-inflated Beta distribution. Likewise, the probability mass of α is assigned to observations of 1 when the one-inflated Beta distribution (BEOI) is assumed instead, namely, $y \sim \mathrm{BEOI}\left(\alpha,\mu,\phi\right)$. The BEZI and BEOI distributions were implemented in the R package “gamlss.dist” (Ospina 2006).

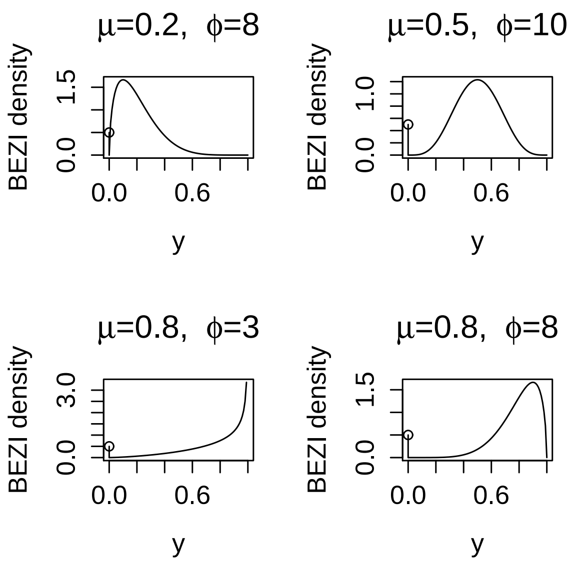


Fig. 1.3: Zero-inflated Beta density plots with $\alpha=0.5$.

Modifications are needed, however, when the response variable is measured on the closed interval [0,1], indicating that the probability masses of both 0 and 1 are beyond that allowed by the underlying Beta distribution. The zero-and-one inflated distribution is a mixture consisting of a continuous distribution on (0,1) and Bernoulli distributions for the values 0 and 1. The distribution function of the mixed discrete-continuous distribution as in Ospina and Ferrari (2010) is given by

$$F_{BEINF}(y;\mu,\phi,\alpha,\gamma)=\begin{cases}0 & y<0\\ \alpha(1-\gamma) & y=0\\ \alpha(1-\gamma)+(1-\alpha)\int_0^y \frac{\Gamma(\varphi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}t^{\mu\phi-1}(1-t)^{(1-\mu)\phi-1}\,dt & y\in(0,1)\\ 1 & y\geq 1\end{cases}$$

We write the mixed discrete-continuous probability function as [1]

$$f_{BEINF}(y;\mu,\phi,\alpha,\gamma)=\begin{cases}\alpha(1-\gamma) & y=0\\ (1-\alpha)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} & y\in(0,1)\\ \alpha\gamma & y=1\end{cases} \tag{1.10}$$

where $0<\alpha,\gamma,\mu<1$, $\phi>0$ and the observations 0 and 1 are assigned the probability mass $\alpha(1-\gamma)$ and $\alpha\gamma$ respectively.

In addition, Ospina and Ferrari (2010) were able to demonstrate that "the zero- and one-inflated Beta distributions are three-parameter exponential family distributions of full rank" and "the zero-and-one-inflated Beta distribution is a four-parameter exponential family distribution of full rank", indicating that future work of regression analysis based on inflated Beta distribution can adopt ideas of GLM. It is noteworthy that for BEINF, there is a more interpretable parametrization in applications, especially when the probability masses of 0 and 1 are associated with independent regressors and of interest in terms of inference. We write the mixed discrete-continuous probability function as

$$f_{BEINF}(y;\mu,\phi,p_0,p_1)=\begin{cases}p_0 & y=0\\ (1-p_0-p_1)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} & y\in(0,1)\\ p_1 & y=1\end{cases} \tag{1.11}$$

This is an appealing parametrization, as the zero-and-one inflated probabilities are explicitly p_0 and p_1, respectively. However, parametrization in (1.11) introduces the constraint $0<p_0+p_1<1$, which leads to the fact that the components p_0 and p_1 in the score vector are correlated. Therefore p_0 and p_1 are not orthogonal, causing a problem in maximum likelihood estimation. In contrast, the parametrization in (1.10) is free of this problem where the components α and γ in the score vector are orthogonal (Ospina and Ferrari 2010).

It is often of interest to model observations on intervals [0,1),(0,1] and [0,1] with regression structures. Based on BEZI and BEOI, Ospina and Ferrari (2012) discussed

[1] For simplicity, only the mixed discrete-continuous probability function will be displayed in the rest of the study.

"a general class of zero-or-one inflated Beta regression models" where only either zero or one is present and the fitting was implemented in "gamlss". With the same density function illustrated as in (1.9), regression structures can be specified with respect to each parameter indexing the distribution as

$$\begin{aligned} g_1(\alpha) &= \eta_1 = f_1(v, \rho) \ , \\ g_2(\mu) &= \eta_2 = f_2(x, \beta) \ , \\ g_3(\phi) &= \eta_3 = f_3(z, \gamma) \ , \end{aligned}$$

where v, x and z are vectors of known covariates, of which vectors of unknown coefficients are denoted by ρ, β and γ. $g(\cdot)$ are suitable link functions. The logit link function is commonly used for μ and α and log link for ϕ. Choices of $f(\cdot)$ include parametric linear and non-linear terms, non-parametric smoothing functions and random effects. They did not, nevertheless, follow up with zero-and-one inflated Beta regression, which is of much importance in applications.

In comparison, Rigby and Stasinopoulos (2005) introduced "gamlss", a modeling framework in which many of the limitations of the generalized linear model (GLM) and generalized additive model (GAM) are overcome. As a general class of univariate regression, "gamlss" provides large flexibility of modeling, partly in that it relaxes the assumption of the exponential family distribution. It has included a variety of distributions of discrete, continuous or mixed distributions, for example, Box-Cox t-distribution, zero-adjusted inverse Gaussian, or skew exponential power distributions, inflated Beta distributions, to name a few. Moreover, it allows not only modeling of the mean (μ) and/or dispersion (σ) but also all other parameters indexing the distribution of the response y, including degrees of freedom of a t-distribution and inflation probabilities. Distributions may have location, scale, skewness and kurtosis parameters. Two shape parameters ν and τ related with skewness and kurtosis suffice for many families of distributions. Modeling each of the four parameters μ, σ, ν, τ gives rise to the following systematic components

$$\begin{aligned} g_1(\mu) &= \eta_1 = f_1(x, \beta) \ , \\ g_2(\phi) &= \eta_2 = f_2(z, \gamma) \ , \\ g_3(\nu) &= \eta_3 = f_3(h, \lambda) \ , \\ g_4(\tau) &= \eta_4 = f_4(k, \rho) \ , \end{aligned}$$

where x, z, h and k are vectors of known covariates and β, γ, λ and ρ are vectors of coefficients of those covariates. Again, $g(\cdot)$ are proper link functions. The function $f(\cdot)$ can be parametric linear, parametric non-linear, non-parametric smoothing terms and random effects, or combinations of these. The user manual of "gamlss" gives a complete list of parametric distributions and applicable functions in regression analysis (Rigby and Stasinopoulos 2009).

It is noteworthy that the model fitting in "gamlss" is evaluated using normalized quantile residuals (or normalized randomized quantile residuals) when the response variable is continuous (or discrete). In cases when y assumes a continuous distribution, the normalized quantile residuals are defined as

$$\hat{r} = \Phi^{-1}(u) \ ,$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution and $u = F\left(y \middle| \hat{\theta}\right)$ is the cumulative distribution function of y (Stasinopoulos, Rigby, and Akantziliotou 2006). The variable u is uniformly distributed on the standard unit interval and therefore the normalized quantile residuals r assume the standard normal distribution. For discrete y, a more general definition is used. For every observation y_i, let (Dunn and Smyth 1996):

$$a_i = \lim_{y \uparrow y_i} F\left(y \middle| \hat{\theta}^i\right),$$
$$b_i = F\left(y_i \middle| \hat{\theta}^i\right) \ .$$

Generate u_i as uniform on $(a_i, b_i]$, then $\widehat{r}_i = \Phi^{-1}(u_i)$. The normalized randomized quantile residuals assume the standard normal distribution.

In "gamlss", the parametrization of zero-and-one inflated Beta distribution is different from that of Ospina and Ferrari's work as in (1.10). It is based on the following parametrization of the Beta distribution:

$$f\left(y;\mu,\sigma^2\right) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\frac{\mu(1-\sigma^2)}{\sigma^2}\right)\Gamma\left(\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)} y^{\frac{\mu\left(1-\sigma^2\right)}{\sigma^2}-1}(1-y)^{\frac{(1-\mu)\left(1-\sigma^2\right)}{\sigma^2}-1} \ ,$$

having $0 < \mu < 1$ and $0 < \sigma < 1$. The mean and variance are $E(y) = \mu$ and $Var(y) = \sigma^2\mu(1-\mu)$. In addition, the parameters μ and σ are connected with the original parameters α and β as in (1.3), with relations $\alpha = \frac{\mu\left(1-\sigma^2\right)}{\sigma^2}$ and $\beta = \frac{(1-\mu)\left(1-\sigma^2\right)}{\sigma^2}$. The mixed discrete-continuous probability function is

$$f_{BEINF}(y;\mu,\sigma,\nu,\tau) = \begin{cases} \frac{\nu}{1+\nu+\tau} & y=0 \\ \frac{1}{1+\nu+\tau}\cdot\frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\frac{\mu(1-\sigma^2)}{\sigma^2}\right)\Gamma\left(\frac{(1-\mu)(1-\sigma^2)}{\sigma^2}\right)} y^{\frac{\mu\left(1-\sigma^2\right)}{\sigma^2}-1}(1-y)^{\frac{(1-\mu)\left(1-\sigma^2\right)}{\sigma^2}-1} & y\in(0,1) \\ \frac{\tau}{1+\nu+\tau} & y=1 \end{cases} \tag{1.12}$$

where the density is indexed by the four parameters μ, σ, ν and τ, with $\nu > 0$ and $\tau > 0$. The probability masses of measures 0 and 1 are associated with the two shape parameters ν and τ through the relations $\nu = p_0/p_2$ and $\tau = p_1/p_2$ with $p_2 = 1 - p_0 - p_1$, where p_0 and p_1 are defined as in (1.11). This parametrization is not ideal in that the interpretation of ν and τ is not straightforward. The parameters μ, σ, ν and τ are modeled with regression structures, rather than μ, σ, p_0 and p_1. It makes difficult the inferences of influences of covariates on p_0, p_1 and/or the overall mean response on the interval [0,1], which may be of interest in applications. That is, its feasibility is at the cost of interpretability.

Two references apply approaches that could address the aforementioned issue but with limitations. Wieczorek and Hawala (2011) employed Bayesian zero-one inflated Beta regression to estimate poverty in U.S counties, in which the parametrization of the underlying BEINF distribution is the ideal. The mean response of the continuous beta distribution μ and the probability masses of 0 and 1 were modeled directly with the same set of regressors and given by

$$\begin{aligned}
\text{logit}\,(\mu) &= x^T\beta_\mu \ , \\
\text{logit}\,(p_0) &= x^T\beta_0 \ , \\
\text{logit}\,(p_1) &= x^T\beta_1 \ .
\end{aligned}$$

Note that the overall mean response is not μ and has the form $E\,(y) = p_1 + (1 - p_0 - p_1)\,\mu$. By specifying prior distributions for the coefficients β_μ, β_0 and β_1, posterior samples were generated using MCMC with Metropolis algorithm, from which inferences could be made for the overall mean $E(y)$, p_0 and p_1 respectively.

A few issues are noteworthy, nevertheless, one of which being the fact that only linear predictors were allowed in systematic modeling. Second, only the mean of the Beta distribution μ and the probability masses p_0, p_1 were modeled with covariates. It can be of concern when the dispersion parameter also changes with covariates. Lastly, the application of Bayesian regression requires appropriate specifications of prior distributions, different choices of which may lead to discrepancy in inference. Improvement was made later in another application, allowing for a random effect in the regression structure of μ. In particular, μ had a stochastic distribution with $\text{logit}(\mu) \sim \text{N}(x^T\beta_\mu, \sigma^2)$ instead of $\text{logit}\,(\mu) = x^T\beta_\mu$ as in the previous application. The analysis was accomplished via the generation of MCMC samples in JAGS from the R software (Wieczorek, Nugent, and Hawala 2012). JAGS ("**J**ust **A**nother **G**ibbs **S**ampling") enables the analysis of Bayesian hierarchal models using MCMC. Limitations of modeling, however, still remain true for the probability masses p_0 and p_1 and the dispersion parameter.

Another approach was presented by Swearingen, Castro and Bursac (2012), the development of which was on the basis of the SAS macro "beta_Regression". It allows for zero-one inflated Beta regression, where probability masses at both 0 and 1 could be modeled with regression structures simultaneously. The mixed discrete-continuous probability function is

$$f_{BEINF}\left(y;\mu,\phi,p_0,p_1\right)=\begin{cases} p_0 & y=0 \\ \left(1-p_0\right)\left(1-p_1\right)\cdot\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}\left(1-y\right)^{(1-\mu)\phi-1} & y\in(0,1) \\ p_1 & y=1 \end{cases} \tag{1.13}$$

where μ and ϕ are parameters indexing the Beta distribution as in (1.8).

Clearly, the probability distribution they specified as in (1.13) is essentially wrong in the sense that it does not sum to one. To date, it still remains an open question of interest to allow for the modeling of p_0 and p_1 directly, with both flexibility and validity.

1.2 Psychometric Functions

In speech perception tests, the speech reception threshold (SRT) is defined as the signal-to-noise ratio (SNR) level at which 50% proportion correct is achieved. SRT is usually modeled with covariates of interest. Typically SRT estimates are obtained by simply averaging certain experimental SNR levels, based on a few established rules of speech perception tests. Psychometric functions can be used alternatively as a curve-fitting technique. The application of psychometric functions is advantageous over averaging rules in that the latter ignores much information about the data, despite its relative ease and simplicity in practice (Dawson, Hersbach, and Swanson 2013).

Psychometric functions, widely used in the practice of psychophysics, describe the relationship between a set of stimuli presented on the x-axis and responses from subjects shown on the y-axis. The response measures are typically in the form of proportion correct, bounded to the standard unit interval. Psychometric functions are applied in order to obtain an estimate of a certain threshold, which is defined according to the research context. In signal detection theory, threshold has two categories: absolute, defined as the level of stimulus at which the observer could identify the existence of the signal; and difference, defined as the smallest difference (d) between the magnitudes of two stimuli enabling the subject to detect the signal (Green, Swets, et al. 1966; McNicol 2005; DeCarlo 1998). Empirical threshold is defined as the intensity of stimulus corresponding to a specific probability value of response, one example of which is the SRT (Treutwein 1995).

Psychometric functions are typically sigmoid with the low asymptote γ, the guessing rate, interpreted as the probability by chance, and the upper asymptote $(1-\lambda)$ with λ being the lapse rate (Klein 2001). The lapse rate λ is the proportion of incorrect responses under optimal signal level, which means that no matter how good the stimulus is, the subject fails to reach 100% correctness. As presented in Wichmann and Hill (2001), the psychometric function is given by

$$\varphi(x;\alpha,\beta,\gamma,\lambda) = \gamma + (1-\lambda-\gamma)\,F(x;\alpha,\beta)\ \ , \tag{1.14}$$

where $F(x;\alpha,\beta)$ is the psychometric function that goes from 0% to 100% and the psychometric function $\varphi(x;\alpha,\beta,\gamma,\lambda)$ represents the data going from γ to $(1-\lambda)$ (Klein 2001). Figure 1.4 gives an example of the psychometric function where the logistic function was used as $F(\cdot)$ and both the lapse rate and the guessing rate were assumed to be 0.1.

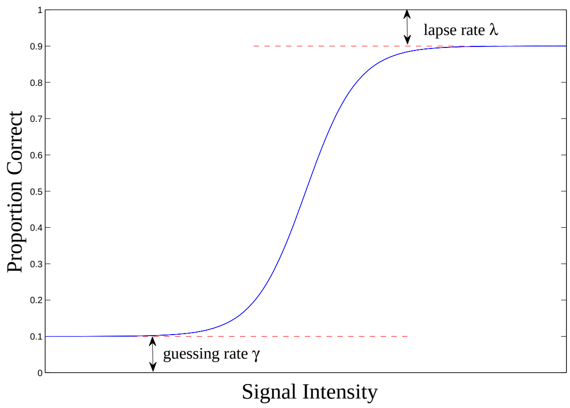


Fig. 1.4: An example of the psychometric function.

According to the documentation of "psignifit", a toolbox with both Python and MATLAB versions available for fitting psychometric functions, the $F(\cdot)$ function can be decomposed into two functions (Fründ, Haenel, and Wichmann 2011). The decomposition is given by

$$F(x;\alpha,\beta) = f(g(x,\alpha,\beta))$$

where $f(\cdot)$ is the "sigmoid" such as the logistic function and $g(\cdot)$ the "core" function. It was stated that "A sigmoid does not have any parameters. Thus, fitting a psychometric function with only a sigmoid would always result in the same psychometric function" (Fründ, Haenel, and Wichmann 2011). The core function is usually a function of the stimulus x indexed by several parameters, with which the combination of a sigmoid enables the psychometric function to assume a variety of shapes determined by those parameters. As an example, the log-core specifies the core function as $g(x,\alpha,\beta) = a\log(x)+b$. Other examples of the core function include Weibull, linear and poly core, to name a few (Fründ, Haenel, and Wichmann 2011). In particular, the psychometric function, with the logistic function as the sigmoid and the ab-core given by $g(x, \alpha, \beta) = \frac{x-a}{b}$, is the standard parametrization in the original version of "psignifit", and is given by

$$F(x; \alpha, \beta) = \frac{1}{1 + \exp\left(-\frac{x-\alpha}{\beta}\right)}$$

where the parameters α and β are related to the threshold and slope of the psychometric function (Dawson, Hersbach, and Swanson 2013).

For cases when subjects suffer from factors independent of stimuli, such as fatigue, the inclusion or modeling of the lapse rate λ is appropriate, or even critical. This can make a substantial difference, as demonstrated by Wichmann and Hill (2001). It is also noteworthy that, although a complete psychometric function is characterized by four parameters including the threshold, slope, lapse rate and guessing rate, it is typically used to estimate a single value, that is, a certain threshold. In addition, the sampling design optimized for a single point estimate is different from that of a complete psychometric function estimation. Treutwein and Strasburger (1999) pointed out that in most adaptive procedures, only one parameter threshold is estimated and the parameters that are not estimated are assigned some reasonable values; for instance, the lapse rate λ is set to zero or a small constant and the guessing rate γ to the expected probability of chance. The context of fitting psychometric functions usually does not involve regression structures where the response is related to a linear combination of a set of covariates, but the statistical estimation of parameters common to data modeling.

There exist several experimental methods for measuring psychometric functions. One popular experimental design is two-alternative forced choice (2AFC) task, in which two spatial or temporal alternatives are presented, with one associated with the test stimulus. For example, in a detection task, for each trial, participants are presented with two time intervals, only one of which contains a signal assigned randomly (Ulrich and Miller 2004). The subject responds by indicating whether the signal appeared in the first or the second interval and the response will be marked as either correct or incorrect. The proportion of correct responses is a measurement of participants' performance. It can be generalized to mAFC, in which observers have m alternatives to choose from and it is appropriate for the guessing rate γ to assume the value $1/m$. Klein (2001) discussed some limitations of 2AFC, and Kaernbach (2001) discussed an unforced choice method (by introducing an "I don't know" option), which offers improvement over standard 2AFC. Other modifications or transformations based on 2AFC are also available.

In speech perception testing, the hearing performance of subjects is usually evaluated as the proportion of correct words identified. The test materials usually consist of word or sentence lists, of which every individual word can be thought of as coming from an enormous pool of vocabulary. In other words, subjects face an open set of choices, and special alterations are needed in the application of psychometric functions. For instance, Dawson, Hersbach and Swanson (2013) treated the guessing rate γ as 0, since the number of choices could be seen as infinite, but kept the lapse rate λ in psychometric function fitting. It is noteworthy that the psychometric function in speech perception testing is in fact applied as a curve fitting technique, aiming to obtain estimates of SRT. As the response measures, SRT estimates are subsequently used for comparisons of different listening conditions, which typically involve simple statistical approaches such as the two sample t-test or repeated measures ANOVA.

The analysis of psychometric functions assumes a binomial model, in almost every study of psychometric function fitting (Kuss, Jäkel, and Wichmann 2005). The experiment can be viewed as a sequence of Bernoulli trials where either a correct or wrong response is produced in each trial. Each observation is the proportion of correct responses in a given trial and the expected probability of a correct response depends on stimulus intensities. Therefore the inference from such analysis should be based on the underlying assumption. Additionally, psychometric function fitting normally involves the estimation of parameters such as α and β as in (1.14), followed by that of a certain threshold.

Wichmann and Hill (2001) established a framework including fitting the psychometric function, deriving error estimates of the parameters and evaluating the goodness of fit, contributing to the development of “psignifit”. In particular, maximum likelihood estimation was applied to obtain parameter estimates. This was achieved via a nonlinear optimization algorithm. The log-likelihood function was given by

$$\ell\left(\theta; y\right) = \sum_{i=1}^{K} \left(\log \begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} + n_i y_i \log \varphi\left(x_i; \theta\right) + \left(n_i - n_i y_i\right) \log\left(1 - \varphi\left(x_i; \theta\right)\right) \right) ,$$

where $\varphi\left(x_i; \theta\right)$ is defined as in (1.14) and n_i is the total number of trials in one block. y_i is the observed data as the proportion of correct responses in the block and K is the number of blocks (Wichmann and Hill 2001).

They also proposed a set of goodness-of-fit tests, one of which was the assessment of overdispersion with deviance. Frequently data for psychometric function analysis is obtained from adaptive procedures which feature a small number of trials n_i in a block. Accordingly, the evaluation based on the asymptotic property of deviance as a χ^2 distribution may be invalid. An alternative Monte-Carlo based technique was applied instead. Specifically, simulated data sets were generated with parametric bootstrapping and for every sample, a deviance value was obtained. This yields the confidence intervals needed. If the deviance of the original data set exceeds the upper bound of the confidence interval, the fit is determined to be poor. Other assessments involved the use of deviance residuals and the detection of influential observations and outliers. The construction of confidence intervals for parameter estimates was also based on parametric bootstrapping.

Yssaad-Fesselier and Knoblauch (2006) described a generalized nonlinear approach for psychometric function fitting using R functions such as "gnlr" from the package "gnlm" and "gnlmix", "hnlmix" from Lindsey's "repeated" library (Lindsey 2001). Still parametric, its generality, however, lies in the ability to model parameters indexing the psychometric function as in (1.14) with the addition of covariates. In one of the examples discussed, $F(x;\alpha,\beta)$ as in (1.14) is a Weibull function, and the psychometric function is therefore

$$\varphi(x;\alpha,\beta,\gamma,\lambda) = \gamma + (1-\lambda-\gamma)\left(1-\exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right]\right) \ .$$

The parameter α is modeled with covariates as

$$\log(\alpha) = x^T\gamma \ .$$

Adding a regression structure enables the comparisons of psychometric functions for different experimental conditions. Moreover, they discussed the possibility of a beta binomial distribution as one of the alternatives in modeling the likelihood, for cases when the variability of the estimated probabilities exceeds that explained by the binomial distribution assumed (Yssaad-Fesselier and Knoblauch 2006).

In contrast, Kuss, Jäkel and Wichmann (2005) adopted Bayesian techniques to analyze psychometric functions and developed the R package "PsychoFun". The Bayesian inference was based on the generation of MCMC posterior samples. With simulated data sets, several point estimates for parameters were compared including the posterior sample mean, the maximum a posteriori (MAP) estimate, the maximum likelihood (ML) estimate and the constrained ML estimate. Bayesian confidence intervals from posterior samples were also compared with the confidence intervals based on parametric bootstrapping. The results suggested that "Bayesian inference methods are likely to lead to more accurate point estimates and confidence intervals than do boostrap-based techiniques" (Kuss, Jäkel, and Wichmann 2005).

In addition, Zchaluk and Foster (2009) proposed a non-parametric method to fit psychometric functions and established the R packages "polynormF", "SparseM" and "modelfree", as well as a companion package "modelfree" in MATLAB. Before the proposal of a nonparametric alternative, they briefly reviewed a parametric approach based on GLM in the estimation of psychometric functions, in which the linear predictor η was expressed as $\eta(x) = g(\varphi(x))$. Note that this approach involves a regression structure where the stimulus intensity is a covariate. As usual, $g(\cdot)$ is the link function. The relation between η and the stimulus intensity x is expressed as $\eta(x) = \beta_0 + \beta_1 x$. The log-likelihood is accordingly given by

$$\begin{aligned}\ell(\beta_0, \beta_1; y) &= \sum_{i=1}^{K} \left(\log \begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} + n_i y_i \log \varphi(x_i) + (n_i - n_i y_i) \log (1 - \varphi(x_i)) \right) \\ &= \sum_{i=1}^{K} \begin{pmatrix} \log \begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} + n_i y_i \log g^{-1}(\beta_0 + \beta_1 x_i) \\ + (n_i - n_i y_i) \log (1 - g^{-1}(\beta_0 + \beta_1 x_i)) \end{pmatrix} ,\end{aligned}$$

where β_0 and β_1 can be estimated by iteratively weighted least squares. The estimate of the psychometric function is therefore given by $\varphi(x) = g^{-1}(\beta_0 + \beta_1 x)$ (Zchaluk and Foster 2009).

More importantly, they proposed local linear estimation in fitting psychometric functions, the principle of which was to estimate the linear predictor locally. For any point x_i in the neighborhood of a specific point x, the corresponding linear predictor $\eta(x_i)$ can be approximated by Taylor expansion as

$$\eta(x_i) \approx \eta(x) + (x_i - x) \eta'(x) \ .$$

We can set equalities as

$$\begin{aligned} \beta_0 &= \eta(x) \ , \\ \beta_1 &= \eta'(x) \ , \end{aligned} \tag{1.15}$$

and the coefficients β_0 and β_1 can be estimated by optimizing the local log-likelihood defined as

$$\ell(a_0, a_1; y) = \sum_{i=1}^{K} \begin{pmatrix} \log \begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} + n_i y_i \log g^{-1}(\beta_0 + \beta_1 (x_i - x)) \\ + (n_i - n_i y_i) \log (1 - g^{-1}(\beta_0 + \beta_1 (x_i - x))) \end{pmatrix} w(x, x_i) \ .$$

This can be viewed as a weighted log-likelihood with the weight function $w(x, x_i)$ defined as

$$w(x, x_i) = \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \ ,$$

where K is the kernel and h is the bandwidth which determines the size of the neighborhood around the specific point x. From (1.15), we can estimate η by the estimate of β_0, that is, $\widehat{\eta}(x) = \widehat{\beta}_0$ and accordingly, the psychometric function is estimated as $\varphi(x) = g^{-1}\left(\widehat{\beta}_0\right)$.

It is necessary to underline the two assumptions of independence and constant probability in the binomial model. It is not uncommon that they can be violated because of factors such as learning during the experimental process, contextual effects or unequal difficulty of testing materials in speech perception testing. Fründ, Haenel and Wichmann (2011) investigated in a simulation study the effect of nonstationarity, referring to cases when "the distribution of the observers' responses changes over the run of the experiment". Specifically, they simulated three different observers: binomial observer, learning observer and betabinomial observer. The first type of observers conform to the two assumptions in the binomial model while the latter two do not. The learning effect was simulated by continuously changing the parameters of the psychometric function, and the betabinomial model assumed that the binomial probability of success changed according to a Beta distribution (Fründ, Haenel, and Wichmann 2011). Both maximum likelihood estimation and Bayesian inference were applied in the psychometric function fitting.

They demonstrated that "nonstationarity can result in serious underestimation of credible intervals" and discussed possible diagnostic measures for nonstationarity (Fründ, Haenel, and Wichmann 2011). In particular, the detection of nonstationarity in the maximum likelihood setting was based on the analysis of the deviance, deviance residuals and the correlation between deviance residuals and the sequence of stimulus presentations, as suggested by Wichmann and Hill (2001). In the Bayesian setting, however, "posterior predictive simulation" was adopted where new samples were generated from the posterior predictive distribution. A Bayesian p-value could then be derived, as the proportion of samples with larger deviance than that of the original sample, which suggests how likely the original data set could be from the assumed model (Fründ, Haenel, and Wichmann 2011). In addition, corrective measures were proposed to inflate the underestimated confidence intervals, because of the basic idea that dependencies were present in cases of nonstationarity (Fründ, Haenel, and Wichmann 2011). The inflation of the underestimated confidence intervals was based on an estimate of the number of independent trials.

To the author's best knowledge, there are not many studies about psychometric functions that considered the violation of the two assumptions in the binomial model and the application of alternative approaches in those cases. However, Zchaluk and Foster (2009) commented that the independence assumption could be relaxed by using the quasi-likelihood method, or the binomial distribution adjusted for overdispersion. Fründ, Haenel and Wichmann (2011) also commented on the feasibility of the beta regression as an alternative in cases of nonstationarity. It is of importance to choose an appropriate model in order to generate reliable inferences.

1.3 The Speech Perception Tests

1.3.1 Test Elements

The hearing-impaired can suffer from hearing losses to different extents, with diminished or even deprived ability to understand everyday speech. In comparison with normally hearing people, listeners with hearing loss typically require higher signal-to-noise ratio (SNR) in order to capture conversational meanings. The introduction of cochlear implants has been demonstrated to be successful in diminishing hearing loss and improving speech recognition for implant recipients (Wilson and Dorman 2008). Growing developments in terms of speech coding, cochlear implant technology and speech perception of the recipients have been witnessed (Gifford, Shallop, and Peterson 2008). It is therefore of particular importance to be able to quantify the performances of cochlear implant recipients in order to assess and compare the effectiveness of different strategies, auditory conditions and other factors of interest.

With respect to test attributes, Mackersie (2002) stated that “an ideal speech perception test will be reliable, will be highly sensitive to differences between test conditions and will correlate well with speech perception abilities in the real world”, issues that need to be taken into consideration in the development of a specific test. In addition, the time and ease of clinical test administration are also factors of importance in practice. Choices of test materials, response measures and speech administration need to be made.

Test materials can be monosyllable word lists that assess lexical information in speech, examples of which include Harvard PB 50 lists, CID W-22 lists and Northwestern University Auditory Test No.6 (Olsen, Van Tasell, and Speaks 1997). There are also meaningful sentence materials that contain semantic information, for example, City University of New York (CUNY) sentence lists (Boothroyd, Hanin, and Hnath 1985), Bamford-Kowal-Bench (BKB) sentence lists (Bench, Kowal, and Bamford 1979) and Speech Perception in Noise (SPIN) (Kalikow, Stevens, and Elliott 1977). One list of the original CUNY sentences consists of five sentences of 10 to 14 words that can be lengthy at times and demanding in terms of the recipient’s memory. Conversely, BKB sentences are shorter with four to six words. One list of SPIN is comprised of 50 sentences, 25 of which are low predictability (LP) sentences and 25 high predictability (HP) sentences, where the predictability is determined based on the last word of a sentence. In cases of HP, there exist semantic links between other words of the same sentence and the key word. An example of HP sentences is: "She made the bed with clean **sheets**" (Kalikow, Stevens, and Elliott 1977). An LP sentence is: "The old man discussed the **dive**" (Kalikow, Stevens, and Elliott 1977). Accordingly, scoring for a SPIN list is defined as the proportion of LP or HP words identified correctly.

In comparison with sentence materials, monosyllable word lists consume less test time and carry more ease in test administration. It has been revealed that "word recognition in quiet was the most commonly used speech test (92%), followed by monosyllabic word recognition in noise (35%) and that only 6% of hearing aid dispensers used sentence-length materials to conduct aided speech testing" (Vaillancourt et al 2005). Despite their popularity in clinical practice, Nilsson, Soli and Sullivan (1994) discussed a few drawbacks associated with monosyllable word lists: first, the reliability of threshold measurements using word lists requires that the individual material should be of equal difficulty, which may not be satisfied in practice; second, the limited number of items in word lists may familiarize listeners with test materials, which may induce learning effects, making the response measurements unreliable; lastly, word lists are not as representative of real world communications as sentences, because of the lack of semantic information, context links as well as intonations and the like.

Test materials are usually grouped in the form of lists, with each forming a block, in which one listening condition would be evaluated. For every sentence or monosyllable word material assessed, a score (proportion correct) is usually obtained as the proportion of correctly identified test items. Proportion correct can be based on either words or morphemes. The definition of "morpheme" is the "minimal meaningful language unit". An example illustrative of morphemes is that "jumping has two morphemes, /jump/ and /ing/" (Dawson, Hersbach, and Swanson 2013). When proportion correct is based on the number of words correctly recognized in word lists or sentences, it can be determined on a word-by-word basis or by just scoring key words. In comparison, Dawson, Hersbach and Swanson (2013) argued the advantage of using morphemes in determining proportion correct: "it can increase the number of test items in the pool for scoring without increasing the number of words and administration time in clinical setting, which also gives rise to the reduction of random measurement error in a score, according to sampling theory". However, in addition, they acknowledged one drawback of the morphemic scoring, that is, the increase of items may result in more perceptually dependent items (Dawson, Hersbach, and Swanson 2013).

It is noteworthy that there exists a problem common to proportion correct, that is, the floor and ceiling effect, since percent scores are bounded to [0,100%]. For example, if the performance of a subject is good and near the top, the proportion correct will not be reflective of any subtle change in listening conditions to a significant extent. The same reasoning applies to the situations where the performance is poor and near the floor (Mackersie 2002). To circumvent the limitation, traditionally, the alternative measure SRT is applied in most studies in order to evaluate speech perception. Figure 1.5 illustrates an example of SRT in the psychometric function, defined as the signal intensity (or SNR in the context of speech perception testing) at the threshold criterion of 50% proportion correct. Note that in speech perception testing, 50% is the usual choice of the level for SRT. However, this can be modified depending on the research question of interest.

SRT is computed over a block of monosyllable words or sentences, in speech perception tests that are administered adaptively. That is, the signal or noise level is increased or decreased by a fixed amount depending on the previous response (Nilsson, Soli, and Sullivan 1994). For example, when the response to the current presentation achieves more than 50% correctness, the subsequent presentation level will be decreased by a fixed step size and vice versa. In this fashion, SNR will converge to the corresponding SRT, at an efficient rate. The application of SRT has other advantages such as ease and efficiency in test administration (Vaillancourt et al 2005). It is notable that the determination of the proportion correct through a specific form of scoring decides the direction of the adjustments of the signal or noise level in adaptive testing. The proportion correct over a block of monosyllable words or sentences is later replaced by a single SRT estimate as the response measure.

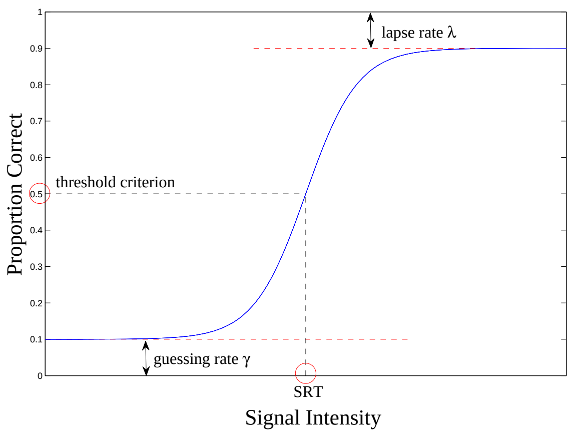


Fig. 1.5: An example of SRT in the psychometric function.

A variety of adaptive tests for many different languages have been developed. Plomp and Mimpen (1979) proposed an adaptive SRT test with Dutch sentences, which were selected, equated for intelligibility and split into 10 lists of 13 sentences. The adaptive procedure was a simple up-and-down procedure, as noted by Levitt and Rabiner (1967), in which the sound levels were either increased or decreased by 2 dB depending on the previous response. SRT was estimated as the average signal level of sentences 5-14 and the test-retest reliability, represented as the standard deviation of individual SRT values was estimated to be 0.9 dB (Plomp and Mimpen 1979). The development of test materials and test protocols in Plomp and Mimpen (1979) laid the foundations for the adaptive tests of many other languages such as Swedish and British English (Vaillancourt et al 2005). In addition, the estimate of the test-retest reliability in their study was often used in comparison with other adaptive tests, which was recommended to not exceed 1.0 dB in order to achieve reasonable test sensitivity (Brand and Kollmeier 2002).

The SPIN test estimates SRT values in American English. The scoring is based on the correct identification of the last word in LP and HP sentences (Kalikow, Stevens, and Elliott 1977). However it was argued that the limited number of available SPIN sentences may restrain the number of conditions that can be tested, and scoring only the last word further reduces the test efficiency since the rest of the sentence is not scored (Nilsson, Soli, and Sullivan 1994). Nilsson, Soli and Sullivan (1994) proposed the Hearing In Noise Test (HINT) with American English sentence lists derived from BKB sentences and equated for sentence difficulties. In the original HINT adaptive procedure, the first four sentences are adjusted with 4 dB step sizes and the remaining sentences with 2 dB step sizes, except for the fifth sentence which was presented at the average of the first five SNR levels (Nilsson, Soli, and Sullivan 1994). For example, suppose the first four sentences are administered at SNR levels [5, 1, 5, 1] and the fourth proportion correct turns out to be greater than 50%. Instead of being -3 after decreasing 4 dB from 1 as expected, the SNR level for the fifth sentence is presented at the average of the first five SNR levels [5, 1, 5, 1, -3], namely, 1.8 dB. This adaptive procedure of HINT was slightly modified in the Australian Sentence Test in Noise (AuSTIN) by excluding the special treatment for the fifth sentence (Dawson, Hersbach, and Swanson 2013).

An alternative paradigm to adaptive testing is to conduct the speech testing at a fixed set of SNR levels, normally used in order to obtain proportion correct as the response. The adaptive SRT method concentrates on the most sensitive region of the performance-intensity function, as opposed to the alternative with fixed SNR levels, a property desirable for detecting subtle differences between listening conditions.

1.3.2 The Analysis of Response Measure

In speech perception tests that are administered at fixed SNR levels, the proportion correct is usually the response of interest. Thornton and Raffin (1978) proposed the modeling of speech-discrimination scores as a binomial variable, since each measure of a subject in response to every word or morpheme, is categorized as correct or incorrect and the reported test score is in fact the proportion of correct responses. Specifically, the test material can be viewed as a pool of stimuli consisting of words or morphemes, each of which can be assigned to two categories, that is, correctly identified or not (Thornton and Raffin 1978).

Implicit in the binomial distribution are two assumptions: (1) that the probability of having a correct response p is fixed for words of uniform difficulty, and (2) that responses are independent of each other. With respect to a specific test form, the application of the binomial distribution needs to be justified or adapted as violations of the two assumptions are not uncommon in practice. For instance, in sentence materials, it is likely that there exist contextual links to some extent between words, which clearly makes the independence assumption questionable. Other probability assumptions, such as the binomial distribution with overdispersion, could be alternatives.

In adaptive testing, typically the response of interest is SRT. There are several approaches that have been used for the estimation of SRT, which can be categorized into approaches that are based on certain averaging rules or on psychometric functions. Dawson, Hersbach and Swanson (2013) adopted four rules to estimate SRT in the adaptive testing, three of which belong to the first category. In particular, the first one is the HINT rule where the SRT is computed as the average of the SNR levels of sentences 5 to 32, or the SNR at which sentence 33 would have been presented depending on the response to sentence 32 (Dawson, Hersbach, and Swanson 2013).

The other two averaging rules estimate SRT as the mean of turns or the mean of 10 turns, where "a turn is defined as a trial in which the adaptive rule changed direction" on the basis of whether a score is greater or less than 50% (Dawson, Hersbach, and Swanson 2013). It is suggestive that the rules with averaging only take into account partial information of the data. In contrast, the fourth rule with psychometric fit adopts all the information with respect to the proportion of correct morphemes identified for each sentence. In other words, all the data is considered in the psychometric function in order to derive the SRT, which is advantageous over the averaging rules from the statistical point of view (Dawson, Hersbach, and Swanson 2013). The dominance of those averaging rules in the speech perception test is simply attributed to its ease and simplicity in application.

Obtaining SRT estimates is simply a middle step in the analysis of speech perception, and is typically followed by simple statistical methods to generate inference. For example, in a study where comparisons of different noise reduction programs and interfering noise types are of interest, sentence lists are employed as the test materials, with each list comprising 16 sentences. Two lists are used to obtain a single estimate of SRT based on the HINT averaging rule, therefore aggregating the original data set. Alternatively, as discussed above, other averaging rules or psychometric functions are applicable in the estimation of SRT. Subsequently, with SRT estimates being the response measures, two-way repeated measures ANOVA is applied as every subject is tested under each listening condition. If any main effect is shown to be significant, the Newman-Keuls post hoc method is performed to enable sequential comparisons of different factor levels. The assumption of sphericity underlying the repeated measures ANOVA is evaluated by Mauchly's test of sphericity, since each factor consists of three levels. In case of any violation, Greenhouse-Geisser adjustment would be applied to alter the degrees of freedom such that the critical values of the F-statistic are inflated accordingly (Dawson, Mauger, and Hersbach 2011).

There are many studies involving similar statistical analysis, the ease and simplicity of SRT being another reason for its popularity. However, it is notable that inference should only be made to the corresponding threshold criterion, for instance, the typical 50% proportion correct. In addition, the successful application of SRT comparisons implicitly assume that psychometric functions for every listening condition are of similar shapes, and differences between those factors only lie in their locations on the x-axis. In other words, psychometric functions shift horizontally without any crossing over. Comparisons based on the single threshold point are thus able to differentiate varying listening conditions. In situations in which this method reveals no such significance for the point corresponding to the 50% proportion correct, it is possible that significant differences do exist at other points or intervals on the S-shaped psychometric function.

2. THE HEARING DATA

2.1 Introduction

Data sets involved in the present analysis are provided by the Cochlear company and obtained from two separate studies which employed the adaptive speech perception test in noise. Each study involves a raw data set in which the response measures are the proportion correct and a derived data set in which SRT estimates are the response measure, derived from the raw data. The two studies adopted a repeated measures design, in which every subject was assessed with each of the listening conditions representing all of the combinations of factors of interest. Randomization of test conditions were taken into account in the procedures. Administered with computer software designed by the Cochlear company, both tests used sentence materials specifically developed for Australians. It is noteworthy that, similar to other established speech perception tests of different languages, the adaptive Australian sentence test in noise (AuSTIN) sentence materials have been equated with difficulty and validated by test-retest reliability (Dawson, Hersbach, and Swanson 2013).

The morphemic score was determined for each sentence, defined as the proportion of morphemes correctly identified. The choice of morphemic scoring over alternatives such as word scoring, or whole sentence scoring in the two studies, was attributed to the consideration that the test items would be increased without increasing the actual number of words tested or administration time. In addition, according to sampling theory, the random measurement error in a score would be reduced by increasing the test items involved. A special rule was applied that, as pointed out by Dawson, Mauger and Hersbach (2011), "in accordance with the HINT rules for subject responses, contracted production of words, and alternative words for articles (e.g., /a/ can be replaced with /the/) and verb tense (e.g., /is/ can be replaced with /was/) were acceptable".

In the protocol, the "signal" or "target" speech, usually presented from the front of subjects, was recorded by a female speaker and the "noise" or "interferer" was a recording of another person, either male or female, which the subjects were instructed to ignore. In both studies, sentences were presented at a fixed 65 dB sound pressure level (SPL), whereas the competing noise level was adjusted adaptively depending on the previous response. More specifically, if a subject scored less than 50% morphemes correct, the interfering noise level for the subsequent sentence would be decreased and vice versa. The corresponding SNRs were recorded as a measure of stimulus intensity. Note that in the two original data sets where morphemic scores were recorded as the response measure, SNR was a covariate of interest as well. However, in the two derived data sets in which SRT estimates were the response measure, SNR was no longer a covariate, because SRT was either estimated as the average of certain SNR levels, or as the SNR level corresponding to 50% correctness from the psychometric function.

2.2 Study One

The first data set is from a study involving seven bilateral Cochlear implant recipients. The test was performed in a way such that every sentence list, comprised of 20 sentences, was evaluated under one listening condition, corresponding to one combination of factors involved. The assessment of every listening condition involved four separate lists for each subject. The adaptive procedure used was in accordance with the Australian Sentence Test in Noise (AuSTIN), that is, the interfering noise level was adjusted by a 4 dB step size for the initial four sentences and a 2 dB step size for the remaining sentences in a list, according to the previous response (Dawson, Hersbach, and Swanson 2013). It is notable that the adaptive protocol was carried out once in every sentence list. In the derived data set, for one list, a single SRT was calculated as the mean of SNR levels of the final 16 sentences, ruling out the initial four sentences. As a result, 20 separate morphemic scores from a list in the original data set aggregated to one SRT value in the derived data set.

Listening conditions being investigated in this study were combinations of three factors of interest: the sound coding strategy ("strategy"), the presenting direction ("noise direction") and the gender of the speaker for the interfering noise ("noise gender"). In particular, there are three different sound coding strategies: A, which is the standard strategy; B and C, as two new strategies. For the interfering noise, it can be presented either from the front (0 degrees) or from the side (90 degrees), with recordings either by a female speaker or a male. There are ten columns in the original data set, labeled as:

(1) subject: the subject identifier from "B1" to "B7".

(2) strategy: the sound coding strategy used including "A", "B" or "C".

(3) date_time: the date and time for each sentence tested.

(4) block: the block number. Each block is in fact one sentence list, which is not in terms of the typical statistical definition but simply an index for the number of lists.

(5) noise_dir: the direction that the noise came from, 0 degree or 90 degrees.

(6) noise_gender: the gender of the interfering talker, "male" or "female".

(7) speech_level: the presentation level of the target speech, which is set fixed at 65 dB SPL.

(8) SNR: the signal-to-noise ratio. It is adjusted within one sentence list adaptively.

(9) num_correct: the number of morphemes correctly recognized in the sentence.

(10) num_items: the total number of morphemes in the sentence

Correspondingly, there are seven columns in the derived data set, that is, subject, strategy, date_time, noise_dir, noise_gender, speech_level and SRT. Four research questions are of importance and interest: (1) What is the best sound coding strategy? are the two new strategies better than the standard one? (2) Is the performance of subjects better with a male interferer or a female interferer? If subject performances vary significantly in the two cases, it implies that subjects are able to detect pitch differences of voice to differentiate the gender of the talker for the competing noise; (3) Is the performance of subjects better when the target speech and noise interferer come from the same direction (0 degree) or different directions (90 degrees)? If the performance is better under the 90 degrees situation, it suggests the existence of a "bilateral benefit", that is, a benefit from listening with two ears in comparison with one ear; (4) Are there any interactions between the three factors?

Figure 2.1a shows the histogram of the proportion correct from the original data set, which clearly resembles a "U" shape. Note the two spikes represent the proportion correct being either zero or one. Figures 2.1b, 2.1c and 2.1d present the histograms of the proportion correct grouped by the three covariates respectively. The distribution of the proportion correct appears to be similar for every factor level. Figures 2.2a, 2.2b and 2.2c are boxplots of SRT grouped by strategy, noise gender and noise direction individually. It seems that for noise gender, the male noise interferer has lower median SRT than female, suggesting that lower SNR, or stimulus intensity, is needed to achieve the 50% proportion correct in the case of the male competing noise. The trend is not very obvious with respect to strategy and noise direction. Further analysis in more depth will be introduced in section 3.2. Note that lower SRT implies better speech perception.

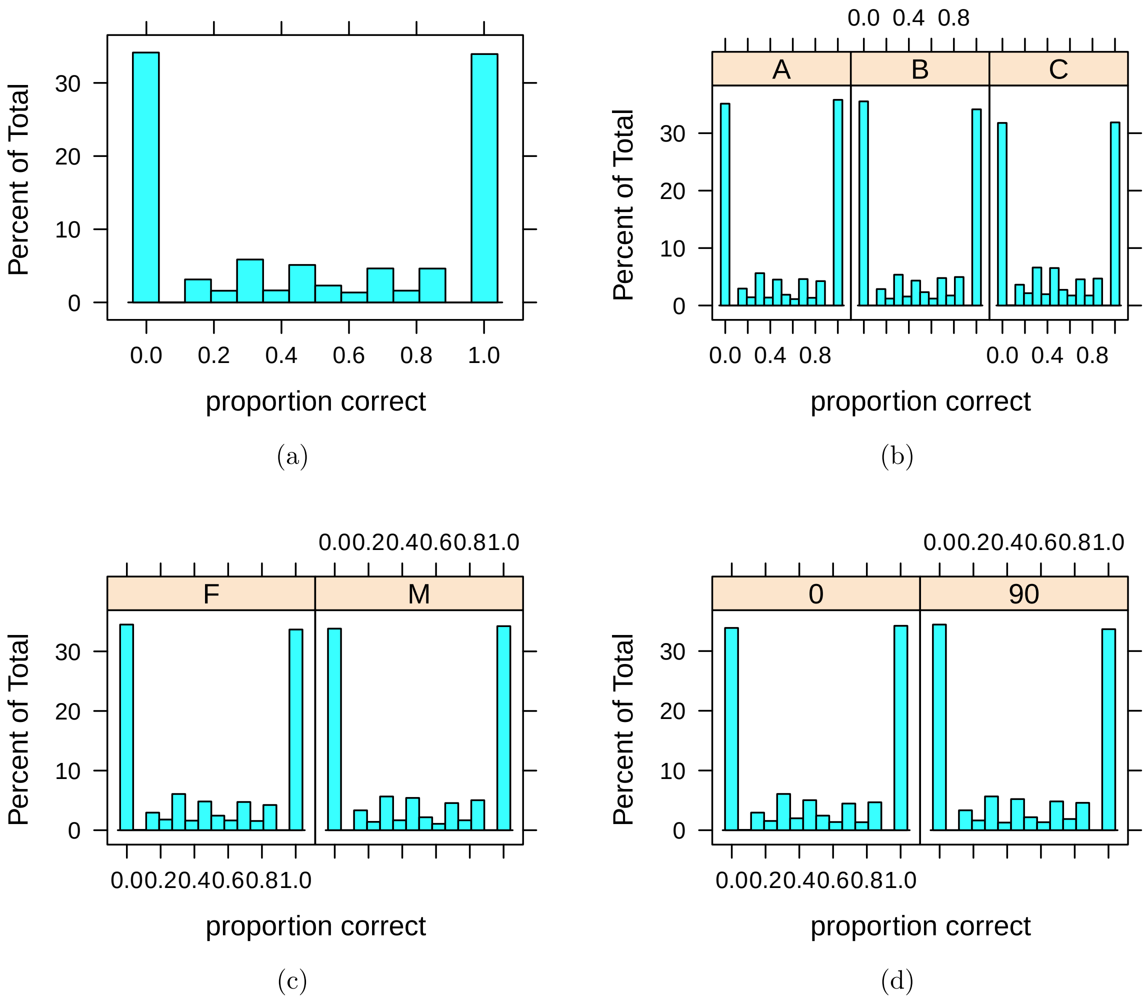
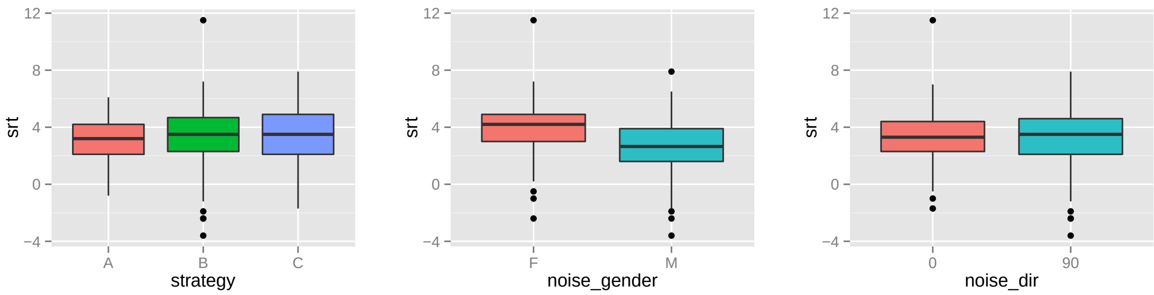


Fig. 2.1: Histogram of the proportion correct: (a) for all observations; (b) grouped by strategy; (c) grouped by noise gender; (d) grouped by noise direction.

2.3 Study Two

The second data set is from a study involving twelve Cochlear implant recipients. This test was also performed in such a way that every sentence list, comprised of 20 sentences, was evaluated under one listening condition, corresponding to one combination of factors involved. However, the assessment of every listening condition now involved two separate lists, for each subject. The original HINT adaptive procedure was adopted. That is, in one list, the adjustment of the interfering noise was a 4 dB step size for the first four sentences and a 2 dB step size for the remaining sentences. The fifth sentence was an exception, for which the corresponding SNR was set as the average of the SNR levels of the initial four sentences and the SNR at which the fifth sentence would have been presented in response to the morphemes correct of the fourth sentence (Nilsson, Soli, and Sullivan 1994). Note that the adaptive protocol was also applied once to every sentence list.



(a) (b) (c)

Fig. 2.2: Boxplots of SRT grouped by: (a) strategy; (b) noise gender; (c) noise direction.

The test conditions were just one factor, that is, the noise reduction algorithm. This was the standard noise reduction algorithm ("Beam") and five variants of the spatial noise reduction algorithm ("SpatialNR (S,0)", "SpatialNR (Z,-3)", "SpatialNR (Z,0)", "SpatialNR (Z,+3)", "SpatialNR (Z,+6)"). There are seven columns in the original data set, labeled as:

(1) subject: the subject identifier from "S1" to "S12".

(2) date_time: the date and time for each sentence tested.

(3) block: the block number. Again, each block is in fact one sentence list, which is not in terms of the typical statistical definition but simply an index for the number of sentence lists.

(4) treatment: noise reduction algorithms.

(5) SNR: the signal-to-noise ratio. It is adjusted within one sentence list adaptively.

(6) num_correct: the number of morphemes correctly recognized in the sentence.

(7) num_items: the total number of morphemes in the sentence.

Correspondingly, there are four columns in the derived data set, consisting of subject, date_time, treatment and SRT. In the second study, the main research questions are: (1) Is the spatial noise reduction technique better than the standard approach "Beam"?; (2) Which variant of the spatial noise reduction is the best?

Figure 2.3a shows the histogram of the proportion correct for the second data set, which also resembles a strong "U" shape. Figure 2.3b shows the histograms of the proportion correct grouped by the noise reduction algorithm. The distribution of the proportion correct appears to be similar for all six noise reduction algorithms. Figure 2.4 gives boxplots of SRT in the derived data set, grouped by noise reduction algorithm. It is suggestive that the standard approach ("Beam") has higher median SRT than the rest, whereas the three variants of spatial noise reduction algorithms including "SpatialNR

(Z,0)", "SpatialNR (Z,+3)", "SpatialNR (Z,+6)" have very similar medians. Higher SRT indicates the need of higher SNR, or stimulus intensity in order to achieve the 50% proportion correct, i.e. worse hearing. Further analysis will be described in section 3.2.

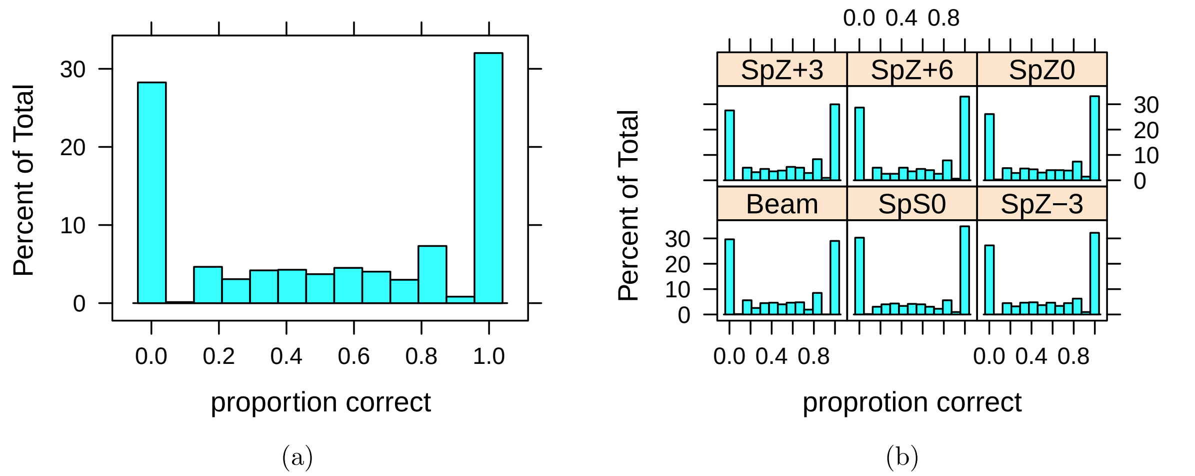


Fig. 2.3: Histogram of the proportion correct: (a) for all observations; (b) grouped by treatments.

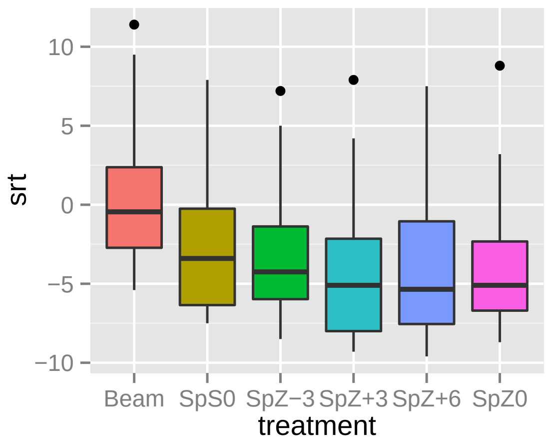


Fig. 2.4: Boxplots of SRT grouped by noise reduction algorithm.

3. MODELING APPROACHES

3.1 Introduction

As introduced in the previous section, raw data sets are obtained directly from hearing tests where morphemic scores are the response measure. In the derived data sets, SRT estimates are the response. Below are two tables that are illustrative of the two types of data respectively. In particular, Table 3.1 involves two sentence lists of 20 sentences each, characterizing two listening conditions. Note that the ratio of "num_correct" and "num_items" gives rise to morphemic scores as proportion correct. Recall that the adaptive testing rule, where the intensity of the present signal or noise level is adjusted based on the previous subject response, is applied to every sentence list, where one SRT estimate would be derived. Correspondingly, the forty observations in Table 3.1 are aggregated to two SRT estimates, given in Table 3.2. The traditional approach to evaluate speech intelligibility of subjects relies on the analysis of SRT estimates, involving simple statistical methods such as the two-sample t-test or repeated measures ANOVA. One apparent drawback of the traditional approach is the loss of information of data, accompanied by reduced statistical power. Considering that the distribution of the proportion correct resembles a "U" shape with two big spikes at values 0 and 1, we propose a zero-and-one inflated beta regression model for the proportion correct, as the new approach of assessing speech intelligibility.

3.2 Traditional Approach

In the two data sets, the morphemic proportion correct is the response measure, but it is argued that the floor and ceiling effect inherent in percent intelligibility scores poses a problem in the comparisons of different listening conditions. The traditional approach to address this issue involves the estimation of SRTs from the original data, followed by the application of some simple statistical methods. It concentrates typically on the point of 50% proportion correct of psychometric function, believed to be sensitive to differences between listening conditions. An implicit assumption underlying the traditional approach is that psychometric functions for every condition bear similar shapes, of which

subject	strategy	date_time	block	noise_dir	noise_gender	speech_level	snr	num_correct	num_items
B1	C	5/3/2013 15:27	1	0	F	65	5	1	7
B1	C	5/3/2013 15:27	1	0	F	65	9	6	6
B1	C	5/3/2013 15:27	1	0	F	65	5	1	7
B1	C	5/3/2013 15:27	1	0	F	65	9	5	5
B1	C	5/3/2013 15:27	1	0	F	65	7	5	5
B1	C	5/3/2013 15:27	1	0	F	65	5	2	7
B1	C	5/3/2013 15:27	1	0	F	65	7	3	5
B1	C	5/3/2013 15:27	1	0	F	65	5	0	6
B1	C	5/3/2013 15:27	1	0	F	65	7	6	6
B1	C	5/3/2013 15:27	1	0	F	65	5	6	6
B1	C	5/3/2013 15:27	1	0	F	65	3	0	7
B1	C	5/3/2013 15:27	1	0	F	65	5	2	6
B1	C	5/3/2013 15:27	1	0	F	65	7	1	6
B1	C	5/3/2013 15:27	1	0	F	65	9	6	6
B1	C	5/3/2013 15:27	1	0	F	65	7	2	6
B1	C	5/3/2013 15:27	1	0	F	65	9	6	6
B1	C	5/3/2013 15:27	1	0	F	65	7	1	7
B1	C	5/3/2013 15:27	1	0	F	65	9	3	6
B1	C	5/3/2013 15:27	1	0	F	65	7	4	6
B1	C	5/3/2013 15:27	1	0	F	65	5	5	6
B1	C	5/3/2013 15:32	2	90	F	65	5	0	6
B1	C	5/3/2013 15:32	2	90	F	65	9	4	4
B1	C	5/3/2013 15:32	2	90	F	65	5	7	7
B1	C	5/3/2013 15:32	2	90	F	65	1	1	7
B1	C	5/3/2013 15:32	2	90	F	65	3	0	5
B1	C	5/3/2013 15:32	2	90	F	65	5	7	8
B1	C	5/3/2013 15:32	2	90	F	65	3	3	6
B1	C	5/3/2013 15:32	2	90	F	65	1	5	7
B1	C	5/3/2013 15:32	2	90	F	65	-1	0	7
B1	C	5/3/2013 15:32	2	90	F	65	1	7	7
B1	C	5/3/2013 15:32	2	90	F	65	-1	0	6
B1	C	5/3/2013 15:32	2	90	F	65	1	1	6
B1	C	5/3/2013 15:32	2	90	F	65	3	4	6
B1	C	5/3/2013 15:32	2	90	F	65	1	3	6
B1	C	5/3/2013 15:32	2	90	F	65	-1	0	7
B1	C	5/3/2013 15:32	2	90	F	65	1	1	6
B1	C	5/3/2013 15:32	2	90	F	65	3	0	5
B1	C	5/3/2013 15:32	2	90	F	65	5	5	6
B1	C	5/3/2013 15:32	2	90	F	65	3	2	5
B1	C	5/3/2013 15:32	2	90	F	65	5	7	7

Tab. 3.1: New approach: an excerpt of the first raw data set (40 observations)

only the locations on the x-axis are reflective of differences between those conditions. The validity of the use of SRT estimates to discriminate different conditions depends on the assumption stated. As illustrations, the two derived data sets, in which SRT estimates are the new response measure, are analyzed in the traditional fashion.

Study One

The first derived data set is comprised of 336 observations, where the factors of interest are "strategy" of 3 levels ("A", "B" and "C"), "noise direction" of 2 levels ("0 degrees" and "90 degrees") and "noise gender" of 2 levels ("female" and "male"). Recall that each subject was evaluated under every listening condition, defined as all combinations of levels of the three factors. Three-way repeated measures ANOVA is applied to SRT estimates to compare different listening conditions and results are given in Table 3.3. The underlying assumption of the repeated measures ANOVA is sphericity, which states that

subject	strategy	date_time	noise_dir	noise_gender	speech_level	SRT
B1	C	5/3/2013 15:27	0	F	65	6.3
B1	C	5/3/2013 15:32	90	F	65	2.1

Tab. 3.2: Traditional approach: an excerpt of the first derived data set with 2 observations, corresponding to the 40 observations in Table 3.1

the variances of differences between all levels of independent factors are the same. If the assumption is violated, inflated F-statistics would be produced as well as unreliable p-values. The validity of the assumption could be evaluated by Mauchly's test of sphericity when repeated measure factors involve three or more levels. The significance level is set at 0.05, if not stated otherwise.

Effect	Type III SS	F value	P value
strategy	4.096	0.543	0.584
noise_gender	165.060	50.490	0.000
noise_dir	0.771	0.179	0.676
strategy×noise_gender	34.576	7.131	0.002
strategy×noise_dir	11.520	2.059	0.137
noise_gender×noise_dir	2.555	1.633	0.212
strategy×noise_dir×noise_gender	1.001	0.380	0.686

Tab. 3.3: Tests of within-subjects effects in the first study

No violation of the sphericity assumption is found in the first analysis concerning all three factors. Accordingly, there is no need for any adjustment. With respect to main effects, "strategy" and "noise direction" appear to be non-significant. The "noise gender" factor, however, is significant, with estimated marginal group means of SRT of 2.59 dB and 3.99 dB for the male and female competing noise respectively. It is notable that larger SRT estimates correspond to higher stimulus intensities required to achieve 50% correct, thus implying worse speech intelligibility. As illustrated by Figure 3.1 where marginal SRT estimates were grouped by strategy and noise gender, it seems that strategies "B" and "C" have lower SRT values with male interferer than with female. It is verified that the interaction term of noise gender and strategy is also significant ($p = 0.002$). The results suggest that strategies under the male competing noise may be associated with better speech perception performances, and subjects may use pitch differences of voice to discern competing talkers. No other interactions terms are found to be significant.

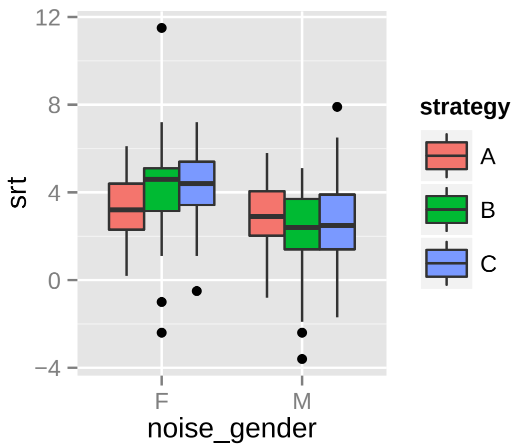


Fig. 3.1: Boxplots of SRT grouped by strategy and noise gender.

The effects of strategy and noise gender on SRT estimates could not be explained simply through the main effects, because of the significance of the interaction term. Instead, simple main effects of strategy are investigated by dividing the data into two subsets, with male and female interfering noise separately. In the first subset where the interfering noise is female, Mauchly's test of sphericity find no sign of violation of the sphericity assumption. The simple main effect of "strategy" becomes significant ($F(2, 54) = 4.676, p = 0.013$), while "noise direction" and other interaction terms are not. Pairwise comparisons with Bonferroni adjustments suggest that in the case of female noise, strategy "A" has significantly lower SRT than "B" ($p = 0.021$), with estimated marginal mean difference of -0.918 dB. No other comparisons are significant. In the second subset with male competing noise, Mauchly's test of sphericity also suggests no violation of the assumption of sphericity. The simple main effect of "strategy", however, is not significant ($F(2, 54) = 1.444, p = 0.245$), as well as "noise direction" and other interaction terms.

In addition, simple main effects of noise gender are also investigated by looking at cases with each strategy individually. Since both "noise gender" and "noise direction" involve only two factor levels, there is no application for the Mauchly's test of sphericity. In the three situations with strategy "A", "B" and "C" respectively, "noise gender" is significant ($F(1, 27) = 5.019, p = 0.033$, $F(1, 27) = 29.096, p < 0.001$, $F(1, 27) = 28.331, p < 0.001$), which suggests that SRT estimates are significantly lower when the interfering noise is male than is female. No other terms are significant.

Study Two

In the second derived data set, there are 144 observations and the factor of interest is "treatment" of 6 levels, that is, the standard noise reduction technique ("Beam") and five variants of the spatial noise reduction technique ("SpatialNR (S,0)", "SpatialNR (Z,-3)", "SpatialNR (Z,0)", "SpatialNR (Z,+3)", "SpatialNR (Z,+6)", abbreviated as "SpS0", "SpZ-3", "SpZ0", "SpZ+3" and "SpZ+6"). Similar to the previous study, each subject was tested with every "treatment" and one-way repeated measure ANOVA is applied to SRT estimates.

Mauchly's test of sphericity suggests that the sphericity assumption was violated marginally ($p = 0.048$) and the Greenhouse-Geisser adjustment is therefore applied to the degrees of freedom of F-statistics involved. It is demonstrated that the "treatment" effect is significant ($F(3.457, 79.506) = 47.401, p < 0.001$). More specifically, the estimated means of SRT are 0.171 dB, -2.583 dB, -3.279 dB, -3.938 dB, -4.446 dB and -3.987 dB for Beam, SpS0, SpZ-3, SpZ0, SpZ+3 and SpZ+6 respectively. Pairwise comparisons with Bonferroni adjustments illustrate that Beam has significantly lower SRT estimates than the other five variants of the spatial noise reduction technique ($p < 0.001$). In addition, SpS0 has significantly smaller SRT estimates than SpZ0, SpZ+3 and SpZ+6 ($p = 0.008, p < 0.001, p = 0.005$ respectively). Pairwise comparisons among the four variants of SpZ-3, SpZ+3 and SpZ+6 suggest that no significant differences exist.

3.3 New Approach

The new approach is based upon the zero-and-one inflated Beta regression, the implementation of which is achieved in the R package"gamlss", in which the location, scale and shape parameters are modeled simultaneously. The "mixed discrete-continuous probability function" of the zero-and-one inflated Beta distribution is given in (1.13).

Study One

In the first study, the raw data set has 6720 observations, in which different listening conditions are comprised of combinations of the three factors ("strategy", "noise gender" and "noise direction"), which are also considered in the covariate structure of zero-and-one inflated Beta regression. The addition of a random effect of "subject" to each modeling is appropriate to account for within-subject correlation. Unlike the traditional analysis where information about SNR is replaced, SNR appears as a covariate in the regression structure of the current study. Interaction terms among the four covariates need to be also considered. With only four covariates, stepwise model selection based on "Generalized Akaike Information Criterion (GAIC)" is adopted for each of the four parameters μ, σ, ν, τ to select possible fixed effects. Results are given in Table 3.4.

The modeling of the first parameter μ, as the expected proportion correct for values between 0 and 1, involves the following selected covariates: "SNR", "noise gender", "strategy" and the interaction of "noise gender" by "strategy". In particular, "SNR" appears to be significant with estimated slope of 0.068, suggesting that higher SNR tends to increase the expected proportion correct for values between 0 and 1. The interaction between "strategy" and "noise gender" is significant, as compared to the reference event of "female" noise by "A" strategy. Results indicate that under male interfering noise, strategies "B" and "C" are associated with higher speech intelligibly for proportion correct on the open interval of (0,1) than "A". It can also be interpreted that for strategies "B" and "C", male competing noise results in better speech perception for proportion correct between 0 and 1 than female.

The selected model for the scale parameter σ involves an intercept only, suggesting that for proportion correct between 0 and 1, the covariates of interest do not affect the shape of the beta distribution. For parameter ν, which is $p_0/(1-p_0-p_1)$, the selected covariates are "SNR", "noise gender", "strategy" and the interaction of "SNR" by "strategy". More specifically, "SNR" is significant and in effect depends on strategy. For all strategies, the effect is negative, indicating that higher SNR tends to decrease ν. The effect of "noise gender" on ν is also significant, with male noise having smaller ν than female noise. Similar analysis also applies to the parameter τ.

Although parameters ν and τ determine the probability masses for proportion correct 0 and 1, the two probabilities p_0 and p_1 are not modeled with regression structures directly, and the effect of the covariates on these probabilities is difficult to interpret. To better illustrate the effects of independent covariates, fitted probabilities for proportion correct 0 ($\hat{p}_0$) and 1 ($\hat{p}_1$) are further derived algebraically and shown in the first two plots of Figure 3.2. It is noted that higher SNR levels (i.e. speech that is easier to hear) are accompanied by lower $\hat{p}_0$ which corresponds to fewer occurrences of proportion correct being 0. Higher SNR levels also tend to produce higher $\hat{p}_1$, illustrative of more occurrences of proportion correct being 1, and accordingly, better speech intelligibility. Note that crossing over of the curves occurs for both $\hat{p}_0$ and $\hat{p}_1$. Fitted means of proportion correct between (0,1), $\hat{\mu}$ are shown in the third plot of Figure 3.2.

Lastly, fitted overall means of proportion correct on [0,1], i.e. $((1-\hat{p}_0-\hat{p}_1)\hat{\mu}+\hat{p}_1))$ are presented in Figure 3.3. It is noteworthy that in the left plot, fitted curves of the three strategies cross over, under both male and female noise, different from the assumption of the traditional approach in which the curves move horizontally along the SNR levels. There are probably two different scenarios with respect to strategy performances depending on SNR levels. In cases when SNR is small, the overall proportion correct is highest with strategy “B”, followed by “C” and “A”, while strategy “A” seems to outperform “B” and “C” once SNR gets larger. It is likely that strategies have relatively varying performances in response to exposures of different SNR levels, situations that the traditional approach could not unveil. It also needs to be pointed out that the random effects of “subject” are taken as 0 when computing fitted probabilities and means.

The diagnostic plots of the fitted model are shown in Figure 3.4, where we can tell that the normality assumption of quantile residuals is satisfied and the fitting of zero-and-one inflated Beta model in the current study is appropriate.

Coefficients		Estimate	Std. Error	t-value	P-value
parameter	μ				
link	logit				
Intercept		-0.184	0.054	-3.406	<0.0001
snr		0.068	0.007	9.580	<0.0001
noise_gender					
F		—	—	—	—
M		-0.020	0.066	-0.305	0.760
strategy					
A		—	—	—	—
B		-0.036	0.066	-0.542	0.588
C		-0.168	0.063	-2.687	0.007
noise_gender×strategy					
M:B		0.184	0.092	1.995	0.046
M:C		0.213	0.089	2.406	0.016
parameter	σ				
link	logit				
Intercept		-0.302	0.020	-15.280	<0.0001
parameter	ν				
link	log				
Intercept		1.129	0.083	13.556	<0.0001
snr		-0.290	0.022	-13.198	<0.0001
noise_gender					
F		—	—	—	—
M		-0.315	0.058	-5.414	<0.0001
strategy					
A		—	—	—	—
B		-0.335	0.102	-3.283	0.001
C		-0.366	0.105	-3.489	<0.001
snr×strategy					
snr:B		0.119	0.028	4.312	<0.0001
snr:C		0.025	0.029	0.871	0.384
parameter	τ				
link	log				
Intercept		-0.866	0.098	-8.819	<0.0001
snr		0.241	0.016	14.790	<0.0001
noise_gender					
F		—	—	—	—
M		0.262	0.132	1.984	0.047
strategy					
A		—	—	—	—
B		-0.179	0.098	-1.815	0.070
C		-0.579	0.010	-5.808	<0.0001
noise_gender×strategy					
M:B		0.132	0.138	0.961	0.336
M:C		0.389	0.138	2.810	0.005
snr×noise_gender					
snr:M		-0.029	0.022	-1.298	0.194

Tab. 3.4: Modeling of μ, σ, ν and τ in the first study

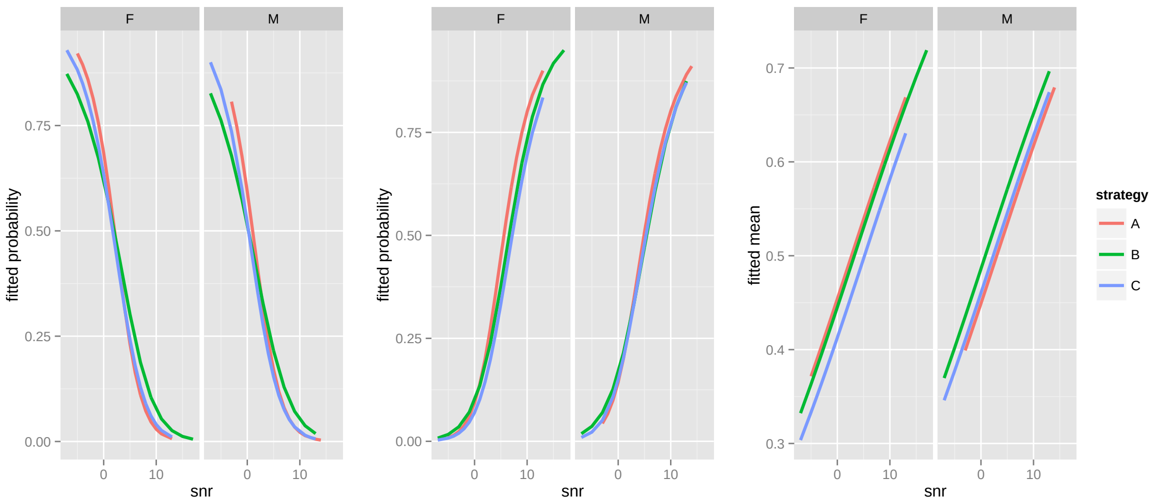


Fig. 3.2: Fitted probabilities for proportion correct 0 and 1, $\hat{p}_0$, $\hat{p}_1$, and fitted means of proportion correct between (0,1), $\hat{\mu}$.

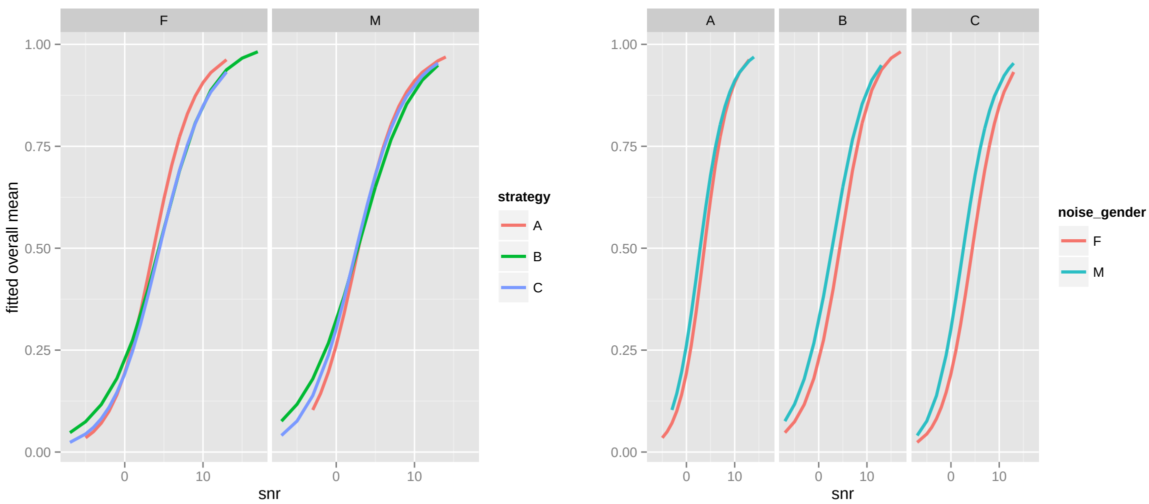


Fig. 3.3: Fitted overall means of proportion correct by SNR, noise gender and strategy.

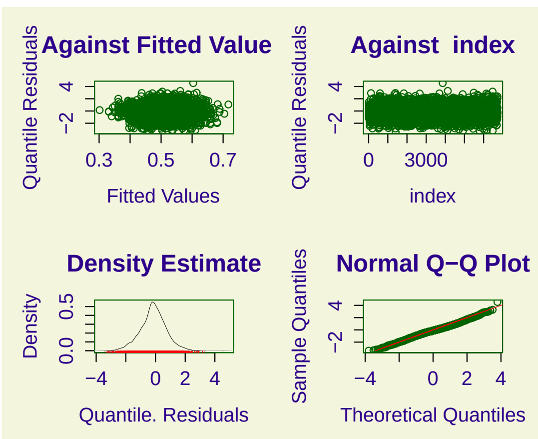


Fig. 3.4: The residual plot for the first study.

It has been demonstrated that “SNR”, “strategy” and “noise gender” can affect the speech intelligibility, of which the effects on the overall proportion correct, however, are pooled over the four separate modeling and only shown visually, not quantitatively. For instance, in the right plot of Figure 3.3, male competing noise tends to produce higher speech intelligibility scores for strategies “B” and “C”, but without quantified magnitudes of difference. Formal hypothesis tests of our research questions would be difficult to formulate and implement, as they go across models for four parameters. Instead, we compute bootstrap confidence intervals for fitted overall means of proportion correct.

More specifically, parametric bootstrap is applied, in which estimates of the four parameters are used to generate pseudo observations and the random effects of subjects are assumed to be zero. Each generated sample contains the same number of observations under every combination of “strategy”, “noise gender” and “SNR”, as in the original data. Every sample is fitted with zero-and-one inflated Beta regression and fitted overall means of proportion correct are obtained. The process is repeated 500 times and confidence intervals of proportion correct are constructed and plotted. Results are presented in Figures 3.5 and 3.6.

Considering that “noise gender” and “strategy” interact, we need to look at differences between levels of one factor by fixing the second factor. The left plot in Figure 3.5 illustrates the constructed confidence intervals of proportion correct grouped by strategy, when the competing noise is female. It seems that the three shaded curves overlap to different extents along SNR levels. At low SNR levels, the three curves almost entirely overlap but start to separate at higher SNR levels, with A being a better strategy. Likewise, the right plot in Figure 3.5 shows the constructed confidence intervals of proportion correct grouped by strategy, when the competing noise is male. Unlike the previous situation, the three curves overlap almost completely at higher SNR levels. Lastly, Figure 3.6 gives the constructed confidence intervals of proportion correct for “noise gender”, under each “strategy” respectively. We can conclude that for strategies “B” and “C”, at a certain range of SNR levels, the male competing noise achieves better speech intelligibility than female, of which the difference is significant.

It is noteworthy that the two studies involve adaptive testing, which is developed to specifically obtain SRT estimates. More observations are therefore at SNR levels near the threshold point, leading to relatively sparse observations at other SNR levels, in particular the extremes of the SNR scale, as shown in Figure 3.7. The proposed new approach works on the full data of proportion correct, instead of derived SRTs. It would benefit from more balanced SNR levels and numbers of observations, the lack of which could lead to reduced ability to discern different listening conditions at the extremes of the SNR scale. Another statistical issue introduced by the adaptive testing design is a lack of independence of observations over time.

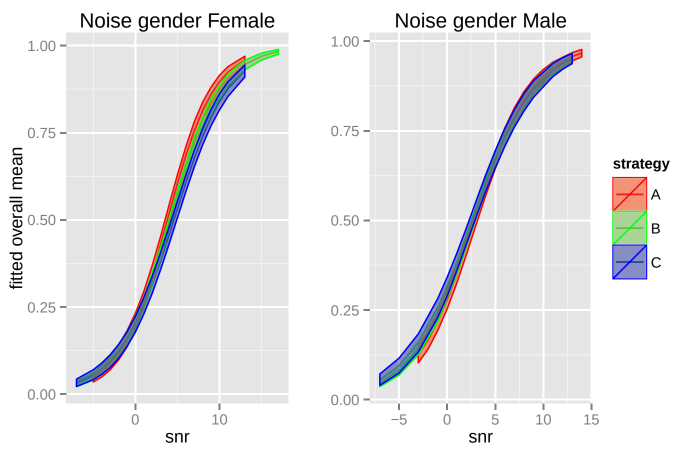


Fig. 3.5: Fitted overall means of proportion correct under strategy and noise gender with 500 bootstrap samples.

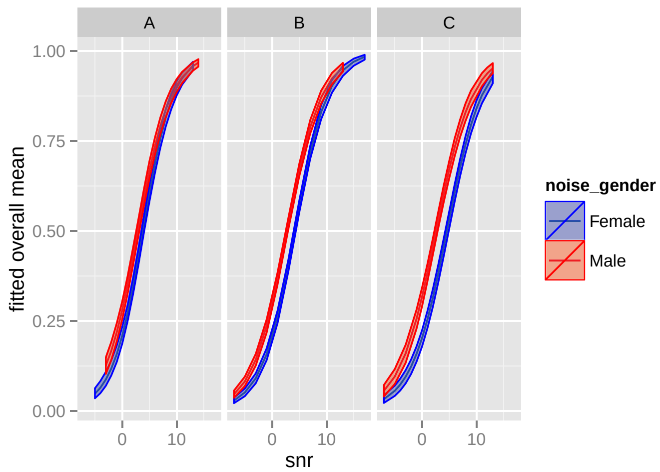


Fig. 3.6: Fitted overall means of proportion correct under strategy and noise gender with 500 bootstrap samples.

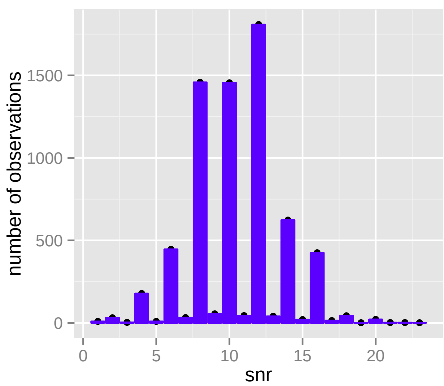


Fig. 3.7: The number of observations at each SNR level.

Study Two

In the second study, the raw data contains 3744 observations, involving only one factor of interest "treatment", consisting of six levels ("Beam", "SpatialNR (S,0)", "SpatialNR (Z,-3)", "SpatialNR (Z,0)", "SpatialNR (Z,+3)", "SpatialNR (Z,+6)", abbreviated as "Beam", "SpS0", "SpZ-3", "SpZ0", "SpZ+3" and "SpZ+6"), which are different noise reduction techniques. Zero-and-one inflated beta regression is applied, in which the four parameters of μ, σ, ν, τ are modeled with "SNR" and "treatment". Summaries of the modeling of μ, σ, ν and τ are given in Table 3.5.

The modeling of σ involves again an intercept only, while that of the other three parameters involves both "SNR" and "treatment". A random effect of "subject" is also considered in each modeling. In addition, no interaction term of "SNR" by "treatment" is needed in the four regression structures, which suggests that differences between the six techniques do not change across different SNR levels. In other words, the sigmoid curves of treatments, where the SNR levels are on the x-axis and the proportion correct on the y-axis, have similar shapes but different locations on the x-axis. It is in agreement with the assumption of the traditional approach.

In the modeling of μ, "SNR" is highly significant with an estimated slope of 0.071, suggesting that higher SNR is linked with larger estimated means of proportion correct between 0 and 1. The differences between the reference group "Beam" and the five variants of spatial noise reduction technique are also significant. In particular, the positivity of the estimated slopes for the five variants indicates higher estimated means of proportion correct between 0 and 1 than the reference group "Beam".

In the modeling of ν, estimated slopes for the five variants are all negative, with p-values much less than 0.05, suggesting their connections with lower ν than "Beam". In contrast, estimated slopes for the five variants are positive in the modeling of τ, which implies the association with higher τ than "Beam". Similar to the first study, fitted probabilities for proportion correct of 0 and 1 are derived through algebraic relations with ν and τ, as shown in Figures 3.8a and 3.8b. As in the first study, the subject random effect is taken as zero. The plots indicates that "Beam" tends to result in higher probability mass for proportion correct of 0 and lower probability mass for proportion correct of 1, than the five variants. That is, the five variants all outperform "Beam", judging simply from the fitted values.

Fitted means of proportion correct between (0,1) is illustrated in Figure 3.8c. Fitted overall means of proportion correct on [0,1] is given in Figure 3.8d. It is obvious that "Beam" has lower expected means for proportion correct between (0,1) than the five variants. The same result applies to proportion correct on [0,1]. Figure 3.9 shows the

Coefficients		Estimate	Std. Error	t-value	P-value
parameter	μ				
link	logit				
Intercept		0.071	0.052	1.369	0.171
snr		0.071	0.005	14.634	<0.0001
treatment					<0.0001
Beam		—	—	—	—
SpS0		0.200	0.078	2.566	0.010
SpZ-3		0.236	0.075	3.125	0.002
SpZ+3		0.391	0.076	5.110	<0.0001
SpZ+6		0.365	0.078	4.705	<0.0001
SpZ0		0.336	0.076	4.396	<0.0001
parameter	σ				
link	logit				
Intercept		-0.333	0.024	-14.040	<0.0001
parameter	ν				
link	log				
Intercept		-0.479	0.097	-4.950	<0.0001
snr		-0.289	0.009	-30.838	<0.0001
treatment					<0.0001
Beam		—	—	—	—
SpS0		-0.560	0.139	-4.042	<0.0001
SpZ-3		-1.028	0.142	-7.260	<0.0001
SpZ+3		-1.388	0.144	-9.623	<0.0001
SpZ+6		-1.170	0.144	-8.148	<0.0001
SpZ0		-1.251	0.143	-8.739	<0.0001
parameter	τ				
link	log				
Intercept		-0.624	0.096	-6.490	<0.0001
snr		0.254	0.009	29.143	<0.0001
treatment					<0.0001
Beam		—	—	—	—
SpS0		1.009	0.136	7.443	<0.0001
SpZ-3		0.943	0.137	6.876	<0.0001
SpZ+3		1.092	0.141	7.752	<0.0001
SpZ+6		1.193	0.139	8.569	<0.0001
SpZ0		1.160	0.138	8.384	<0.0001

Tab. 3.5: Modeling of μ, ν and τ in the second study

diagnostic plots for the fitted model. It is evident that the normality assumption of quantile residuals is satisfied and the zero-and-one inflated Beta model is appropriate. However, the observed differences between the six techniques are illustrated only visually but not quantitatively, the significance of which could not be concluded.

Therefore, the same approach from the first study involving parametric bootstrap is applied in order to construct confidence intervals of proportion correct for different treatments. The result is shown in Figure 3.10, with constructed confidence intervals of proportion correct grouped by techniques. It is clear that for almost every SNR level, "Beam" has significantly lower speech intelligibility scores than the five variants. The shaded curves for SpZ0, SpZ+3 and SpZ+6 nearly overlap each other entirely, suggesting that with the current data, their performances can be viewed as indistinguishable. For a range of SNR levels, it seems that SpS0 has smaller percent scores than SpZ0, SpZ+3 and SpZ+6 but not very evidently.

The second study faces the same problem as in the first one, in which the data is obtained in an adaptive fashion, with less spanned SNR levels. The number of observations are plotted in Figure 3.11, in which individual SNR levels are aggregated to the following intervals: (-14.5, -11.5), (-11.5, -8.5), (-8.5, -5.5), (-5.5, -2.5), (-2.5, 0.5), (0.5, 3.5), (3.5, 6.5), (6.5, 9.5), (9.5, 12.5), (12.5, 15.5), (15.5, 18.5). The differences between the five variants would be sharper and better detected if SNR levels were presented more evenly and not adaptively, with less potential for violations of independence of observations over time.

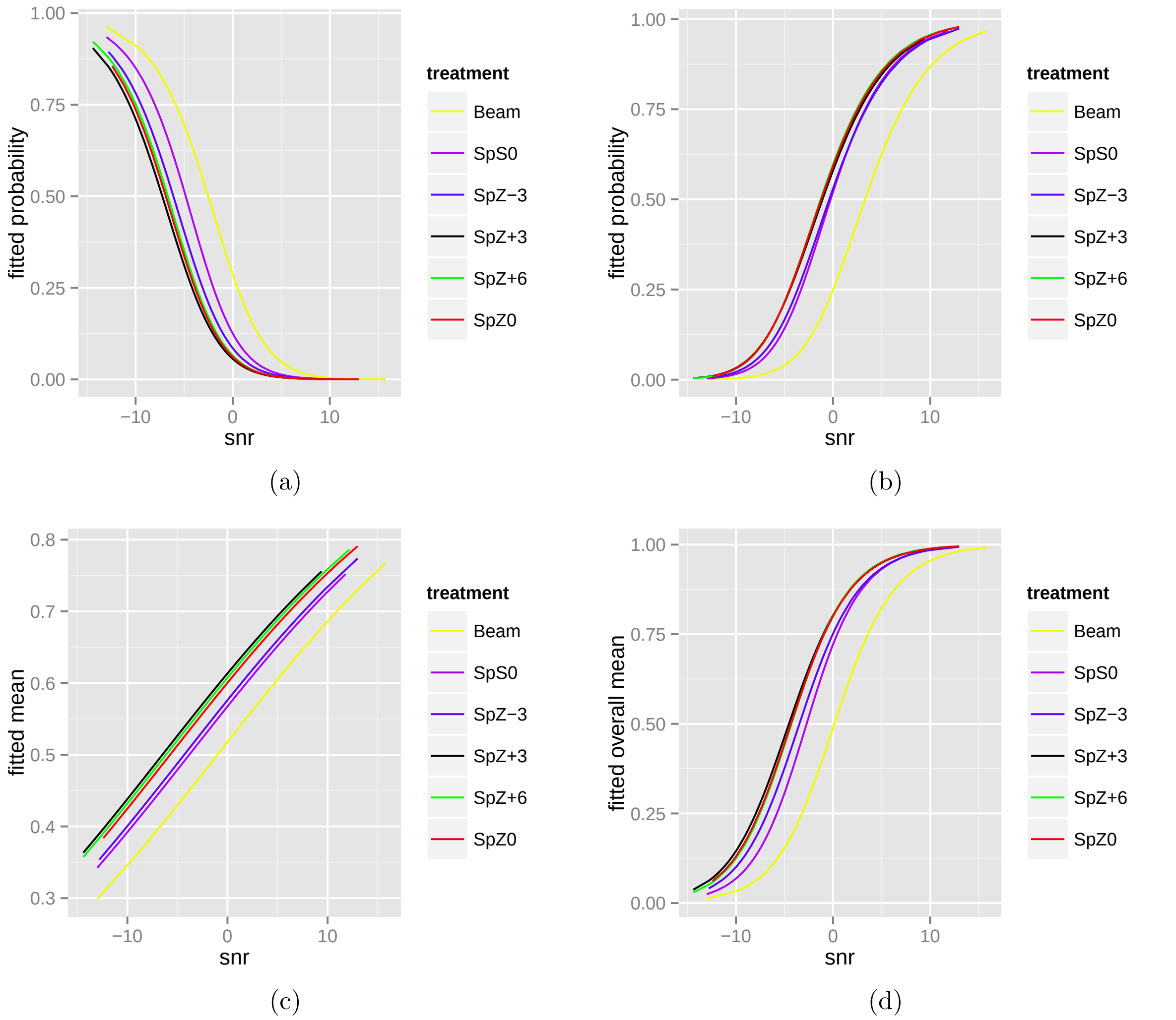


Fig. 3.8: (a) Fitted probability for proportion correct of 0, $\hat{p}_0$; (b) Fitted probability for proportion correct of 1, $\hat{p}_1$; (c) Fitted means of proportion correct between (0,1), $\hat{\mu}$; (d) Fitted overall means of proportion correct by SNR and treatment.

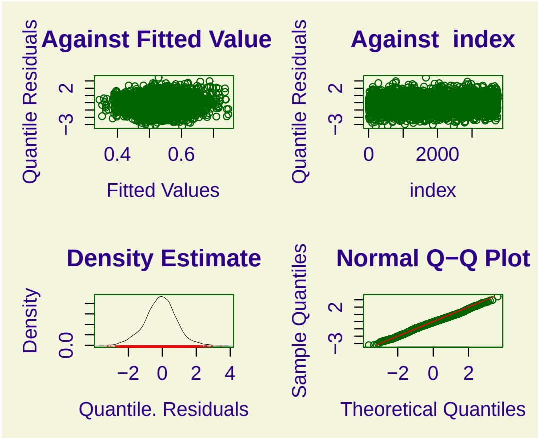


Fig. 3.9: The residual plot for the second study.

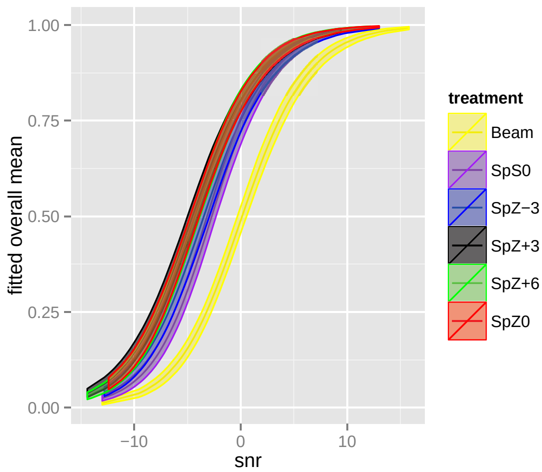


Fig. 3.10: Fitted overall means of proportion correct by treatment with 500 bootstrap samples.

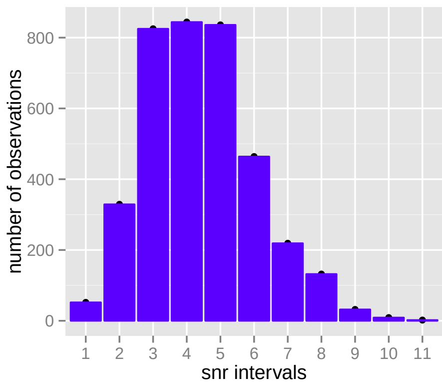


Fig. 3.11: The number of observations at SNR intervals: (-14.5, -11.5), (-11.5, -8.5), (-8.5, -5.5), (-5.5, -2.5), (-2.5, 0.5), (0.5, 3.5), (3.5, 6.5), (6.5, 9.5), (9.5, 12.5), (12.5, 15.5), (15.5, 18.5), labeled as 1 to 11 respectively.

4. DISCUSSION

It is of importance to be able to quantify speech recognition performance. The traditional approach first estimates SRT as the response measure, which are subsequently analyzed by simply statistical methods such as the t-test or the repeated measures ANOVA. However, this approach suffers from a few problems, and we instead propose a zero-and-one inflated Beta regression model, which is advantageous over the traditional one in several aspects.

Firstly, the new approach employs the complete data of proportion correct, whereas the traditional approach relies on aggregated data of SRT. As an illustration, the forty observations of proportion correct in Table 3.1 are replaced by two observations of SRT in Table 3.2. Therefore sample sizes involved in the traditional approach are much reduced. As shown by the first study, a total of 6720 observations are used in the new approach, while only 336 observations in the traditional one. The reduced sample size is associated with decreased power of modeling.

Secondly, the new approach can generate more inference than the single threshold point as in the traditional method, especially at either end of scale or crossover effect of psychometric functions. Study one is a good example of the crossover effect, as in the left plot of Figure 3.5. The new approach reveals that there exist probably two different scenarios of strategy performances, depending on SNR levels. In cases when SNR is small, the overall proportion correct is highest with strategy "B", followed by "C" and "A", while strategy "A" seems to outperform "B" and "C" once SNR gets bigger. The traditional approach, however, fails to identify the crossover effect.

In addition, valid inference based on the threshold point implicitly requires that psychometric functions are of similar shapes for factor levels of interest, differences between which are only reflected by x-axis locations. That is, no crossover happens. Study two is an illustration of where the assumption is true, and the traditional approach can be validly applied. In contrast, the new approach has no such limitation.

Thirdly, psychometric functions, used in the estimation of SRT, assume the binomial distribution for proportion correct. However, the assumptions of independence and constant probability are usually violated because of factors such as contextual effects or unequal difficulty of testing materials. Proportion correct usually resembles a "U" shape, as in

Figure 2.1a, with two big spikes at values 0 and 1. More often those extreme values at 0 and 1 are meaningful and of interest. The new approach is able to accommodate the special distributional properties of proportion correct, which is thus more appropriate.

Lastly, random effects of subjects can be quantified and assessed in the new approach, which is not applicable in the traditional one. Furthermore, it is often of benefit to be able to assess individual performance, as an initial screening. Those initial results offer testers information about the subject or the design. The new approach is found to be also applicable to observations of each individual, while the traditional one fails.

It is noteworthy that the estimation of SRT involves speech perception testing that is administered adaptively, in which the signal or noise level is increased or decreased by a fixed step size based on the previous response (Nilsson, Soli, and Sullivan 1994). That is, when the response to the current presentation achieves more than 50% correctness, the subsequent presentation level will be decreased by a fixed level and vice versa. In this adaptive testing, SNR will converge to the corresponding SRT, an ideal property for the traditional approach. Two issues arises, however, for the application of the new approach to proportion correct obtained in this adaptive manner, one of which being the fact that observations at each experimental SNR level are very unbalanced, as shown in Figures 3.7 and 3.11. One would expect that more observations center around SNR levels near the threshold point, with sparse counts at the two ends of the scale. Another statistical issue introduced by the adaptive testing design is a lack of independence of observations over time, in that SNR levels are not fixed, but adjusted depending on the previous response.

The new approach relies on proportions correct, instead of SRT, which is therefore free of many issues associated with the traditional approach, such as the need of an adaptive testing or the estimation of SRT. It has been shown that data obtained adaptively pose two problems to the application of the new approach. Sharper contrasting differences could be obtained for the current studies, given a design which is more balanced over the range of SNR, and administered in a non-adaptive fashion.

REFERENCES

Bench, J., Å. Kowal, and J. Bamford (1979). The bkb (bamford-kowal-bench) sentence lists for partially-hearing children. *British Journal of Audiology 13*(3), 108–112.

Boothroyd, A., L. Hanin, and T. Hnath (1985). A sentence test of speech perception: Reliability, set equivalence, and short term learning. *New York, NY: City University of New York*.

Bottai, M., B. Cai, and R. E. McKeown (2010). Logistic quantile regression for bounded outcomes. *Statistics in Medicine 29*(2), 309–317.

Brand, T. and B. Kollmeier (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America 111*(6), 2801–2810.

Branscum, A. J., W. O. Johnson, and M. C. Thurmond (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics 49*(3), 287–301.

Carrasco, J. M., S. L. Ferrari, and R. B. Arellano-Valle (2014). Errors-in-variables beta regression models. *Journal of Applied Statistics 41*(7), 1530–1547.

Cohen, A. C. (1963). *Estimation in mixtures of discrete distributions*. Statistical Pub. Society.

Cribari-Neto, F. and K. L. Vasconcellos (2002). Nearly unbiased maximum likelihood estimation for the beta distribution. *Journal of Statistical Computation and Simulation 72*(2), 107–118.

Cribari-Neto, F. and A. Zeileis (2010). Beta regression in R. *Journal of Statistical Software 34*(2), 1–24.

Dawson, P. W., A. A. Hersbach, and B. A. Swanson (2013). An adaptive australian sentence test in noise (austin). *Ear and Hearing 34*(5), 592–600.

Dawson, P. W., S. J. Mauger, and A. A. Hersbach (2011). Clinical evaluation of signal-to-noise ratio–based noise reduction in nucleus® cochlear implant recipients. *Ear and Hearing 32*(3), 382–390.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods 3*(2), 186–205.

Dunn, P. K. and G. K. Smyth (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics 5*(3), 236–244.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*(7), 799–815.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

Frederic, P. and F. Lad (2008). Two moments of the logitnormal distribution. *Communications in Statistics: Simulation and Computation® 37*(7), 1263–1269.

Fründ, I., N. V. Haenel, and F. A. Wichmann (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision 11*(6), 367–

382.

Ghosh, S. K., P. Mukhopadhyay, and J.-C. J. Lu (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference 136*(4), 1360–1375.

Gifford, R. H., J. K. Shallop, and A. M. Peterson (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiology and Neurotology 13*(3), 193–205.

Green, D. M., J. A. Swets, et al. (1966). *Signal detection theory and psychophysics*, Volume 1. Wiley New York.

Grün, B., I. Kosmidis, and A. Zeileis (2012). Extended beta regression in r: Shaken, stirred, mixed, and partitioned. Technical Report 11.

Johnson, S. and N. Kotz (1969). *Discrete distributions*. Houghton Mifflin Company.

Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics 63*(8), 1377–1388.

Kalikow, D., K. N. Stevens, and L. Elliott (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America 61*(5), 1337–1351.

Kieschnick, R. and B. D. McCullough (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling 3*(3), 193–213.

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics 63*(8), 1421–1455.

Kuss, M., F. Jäkel, and F. A. Wichmann (2005). Bayesian inference for psychometric functions. *Journal of Vision 5*(5), 478–492.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Lesaffre, E., D. Rizopoulos, and R. Tsonaka (2007). The logistic transform for bounded outcome scores. *Biostatistics 8*(1), 72–85.

Levitt, H. and L. Rabiner (1967). Use of a sequential strategy in intelligibility testing. *The Journal of the Acoustical Society of America 42*(3), 609–612.

Lindsey, J. (2001). *Nonlinear models in medical statistics*. Oxford University Press.

Mackersie, C. L. (2002). Tests of speech perception abilities. *Current Opinion in Otolaryngology & Head and Neck Surgery 10*(5), 392–397.

Mayr, A., N. Fenske, B. Hofner, T. Kneib, and M. Schmid (2012). Generalized additive models for location, scale and shape for high dimensional dataa flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(3), 403–427.

McNicol, D. (2005). *A primer of signal detection theory*. Psychology Press.

Mu, Y. and X. He (2007). Power transformation toward a linear regression quantile. *Journal of the American Statistical Association 102*(477), 269–279.

Nilsson, M., S. D. Soli, and J. A. Sullivan (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America 95*(2), 1085–1099.

Olsen, W. O., D. J. Van Tasell, and C. E. Speaks (1997). Phoneme and word recognition for words in isolation and in sentences. *Ear and Hearing 18*(3), 175–188.

Ospina, R. (2006). The zero-inflated beta distribution for fitting a gamlss. contribution to r package gamlss. dist: Extra distributions to be used for gamlss modelling.

Ospina, R., F. Cribari-Neto, and K. L. Vasconcellos (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics & Data Analysis 51*(2), 960–981.

Ospina, R. and S. L. Ferrari (2010). Inflated beta distributions. *Statistical Papers 51*(1), 111–126.

Ospina, R. and S. L. Ferrari (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis 56*(6), 1609–1623.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis 9*(4), 325–346.

Plomp, R. and A. Mimpen (1979). Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology 18*(1), 43–52.

Rigby, R. and D. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Rigby, R. and D. Stasinopoulos (2009). A flexible regression approach using gamlss in r. *London Metropolitan University, London.*

Schmid, M., F. Wickler, K. O. Maloney, R. Mitchell, N. Fenske, and A. Mayr (2013). Boosted beta regression. *PLOS ONE 8*(4), e61623.

Simas, A. B., W. Barreto-Souza, and A. V. Rocha (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis 54*(2), 348–366.

Smithson, M. and J. Verkuilen (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods 11*(1), 54–71.

Stasinopoulos, D., R. Rigby, and C. Akantziliotou (2006). Instructions on how to use the gamlss package in r. *Accompanying documentation in the current GAMLSS help files,(see also http://www. gamlss. org/).*

Swearingen, C., M. Castro, and Z. Bursac (2012). Inflated beta regression: Zero, one and everything in between. In *SAS Global Forum*, pp. 325–2012.

Thornton, A. R. and M. J. Raffin (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech, Language, and Hearing Research 21*(3), 507–518.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research 35*(17), 2503–2522.

Treutwein, B. and H. Strasburger (1999). Fitting the psychometric function. *Perception & Psychophysics 61*(1), 87–106.

Ulrich, R. and J. Miller (2004). Threshold estimation in two-alternative forced-choice (2afc) tasks: The spearman-kärber method. *Perception & Psychophysics 66*(3), 517–533.

Vaillancourt, V., C. Laroche, C. Mayer, C. Basque, M. Nali, A. Eriks-Brophy, S. D. Soli, and C. Giguère (2005). Adaptation of the hint (hearing in noise test) for adult canadian francophone populations: Adaptación del hint (prueba de audición en ruido) para poblaciones de adultos canadienses francófonos. *International Journal of Audiology 44*(6), 358–361.

Vasconcellos, K. L. and F. Cribari-Neto (2005). Improved maximum likelihood estimation in a new class of beta regression models. *Brazilian Journal of Probability and Statistics 19*(1), 13–31.

Verkuilen, J. and M. Smithson (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics 37*(1), 82–113.

Wichmann, F. A. and N. J. Hill (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics 63*(8), 1293–1313.

Wieczorek, J. and S. Hawala (2011). A bayesian zero-one inflated beta model for estimating poverty in us counties. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association*.

Wieczorek, J., C. Nugent, and S. Hawala (2012). A bayesian zero-one inflated beta model for small area shrinkage estimation. *JSM Proceedings. Forthcoming*.

Wilson, B. S. and M. F. Dorman (2008). Cochlear implants: a remarkable past and a brilliant future. *Hearing Research 242*(1), 3–21.

Yssaad-Fesselier, R. and K. Knoblauch (2006). Modeling psychometric functions in r. *Behavior Research Methods 38*(1), 28–41.

Zchaluk, K. and D. H. Foster (2009). Model-free estimation of the psychometric function. *Attention, Perception, & Psychophysics 71*(6), 1414–1425.