Is the Effort Justified?

Claudia Czado and Axel Munk *

Abstract

Generalized linear models (GLMs) allow for a wide range of statistical models for regression data. In particular, the logistic model is usually applied for binomial observations. Canonical links for GLM's such as the logit link in the binomial case, are often used because in this case minimal sufficient statistics for the regression parameter exist which allow for simple interpretation of the results. However, in some applications, the overall fit as measured by the p-values of goodness of fit statistics (as the residual deviance) can be improved significantly by the use of a noncanonical link. In this case, the interpretation of the influence of the covariables is more complicated compared to GLM's with canonical link functions. It will be illustrated through simulation that the p-value associated with the common goodness of link tests is not appropriate to quantify the changes to mean response estimates and other quantities of interest when switching to a noncanonical link. In particular, the rate of misspecifications becomes considerably large, when the inverse information value associated with the underlying parametric link model increases. This shows that the classical tests are often too sensitive, in particular, when the number of observations is large. The consideration of a generalized p-value function is proposed instead, which allows the exact quantification of a suitable distance to the canonical model at a controlled error rate. Corresponding tests for validating or discriminating the canonical model can easily performed by means of this function. Finally, it is indicated how this method can be applied to the problem of overdispersion.

Key Words: *generalized linear models, goodness of link tests, logistic regression, link function, parametric links, model validation and discrimination, p-value curve, overdispersion.*

AMS Subject Classification: 62F03, 62F04, 62J12

1 Introduction

Generalized linear models (GLMs) allow for the treatment of regression problems in which the response distribution can be chosen as a one parametric exponential family. This includes normal, binomial, Poisson, gamma and inverse Gaussian responses (see McCullagh & Nelder (1989)) among many others. For this a link function connecting the mean response with the linear predictor has to be chosen. GLM's with canonical links (for definition see McCullagh & Nelder (1989)), such as the logit link in binomial regression, guarantee maximum information and simple interpretation of the regression parameters, because in this case we obtain a linear model for the natural parameter of the underlying exponential family. For example, the logit link allows for a simple representation of the odds, which aids the interpretation of the results. The concavity of the log likelihood guarantees uniqueness of the MLE.

Canonical links, however, do not always provide the best fit available to a given data set. In this case, the link could be misspecified, which can lead to substantial bias in the regression parameter and the

[*] C. Czado is Associate Professor for Statistics at Technische Universität München, Germany, and A. Munk is Assistant Professor at the Ruhr-Universität Bochum, Universitätsstr. 150, 44780 Bochum, Germany. Parts of this paper were written while C. Czado was visiting the Sonderforschungsbereich 386 'Statistische Analyse diskreter Strukturen' at the Ludwig-Maximilians-Universität München, Germany. The authors would like to thank L. Fahrmeir and the LMU München for its hospitality. C. Czado was supported by research grant OGP0089858 of the Natural Sciences and Engineering Research Council of Canada. The work of A. Munk was partially supported by the Deutsche Forschungsgemeinschaft.

approach to guard against such a misspecification, is to embed the canonical link into a wide parametric class of links $\mathcal{F} = \{F(\cdot, \psi), \psi \in \Psi\}$, which includes the canonical link as a special case when $\psi = \psi_*$, say. Many such parametric link classes for binary regression data have been proposed in the literature. Van Montford and Otten (1976), Copenhaver and Mielke (1977), Aranda-Ordaz (1981), Guerrero and Johnson (1982), Morgan (1983) and Whittmore (1983) proposed one-parameter families, while Prentice (1976), Pregibon (1980), Stukel (1988) and Czado (1992) considered two-parameter families. Link functions for the non binary case were studied by Pregibon (1980) and Czado (1992, 97). Section 2 introduces in detail generalized linear models with a parametric link.

In the following, it is assumed that the true underlying link is a member of such a class $\mathcal{F}$. To protect against link misspecification large sample tests such as the likelihood ratio and the score test are recommended (Pregibon (1980, 1982) and McCullagh and Nelder (1989), Chapter 11) to assess, if a different link will lead to a significant improvement in fit. Hence guarding against link misspecification is treated as the testing problem

$$H : \psi = \psi_* \qquad \text{versus} \qquad K : \psi \neq \psi_*. \tag{1}$$

If H cannot be rejected, the additional consideration of the p-value of the pivot statistic is accepted as a sufficient measure for evidence to keep the canonical model. If H is rejected, maximum likelihood estimation (MLE) is required to estimate jointly the link parameter ψ and the regression parameters. However, practitioner prefer using a canonical model since they allow easier interpretation of the model parameters. For example, the logistic model allows interpretation in terms of the odds ratio. Further, the joint estimation increases significantly the computational effort to analyze the data, because a noncanonical link model $\mathcal{F}$ cannot be performed in standard software packages and special software such as macros have to be written. In addition to the special software requirement, the estimation of the link also inflates the variance of the regression parameters, since the link parameter ψ cannot be chosen orthogonal in the sense of Cox and Reid (1987) to the regression parameters (see Taylor (1988) and Czado (1997)). Further, checks for isolated departures from the model (for a general review see Davison and Tsai (1992)) have been developed so far only for fixed link models (see Pregibon (1981) for logistic regression, Lee (1987, 1988) and Williams (1987) for GLM's).

Therefore, the goal is to provide a statistical tool for answering the following two questions:

Q1 When is the effort justified to switch from a canonical link to a noncanonical link in a GLM ? (Model Discrimination)

Q2 How large is the evidence for the canonical model indicated by a large p-value associated with a goodness of link test for (1)? (Model Validation)

First of all, we would like to point out that *both* questions cannot be sufficiently answered by the consideration of the p-value associated to one of the above mentioned tests for (1). This will be illustrated in a simulation study (Appendix A). We mention, that even after diagnostic tools are used in order to protect against outliers or other isolated deviations from the model, this approach cannot be justified. Roughly speaking, this study indicates two systematic errors. On the one hand, when the data are rather noisy the classical tests will not reject $H : \psi = \psi^*$ with large probability although the true mean response (or other parameters of interest) is far apart from the mean response under the canonical link assumption. On the other hand, when the variation of the estimated link is small or the sample size is too large, we find that these tests lead with high probability to a decision in favor of the noncanonical model - although the effort is not justified, i.e. over a wide range of the mean space these links will be the goodness of link *test* for the testing problem (1), rather this is intrinsically related to the misleading *hypothesis* H in (1). Whenever the null hypothesis is rejected, no information about the amount of discrepancy to the noncanonical model is involved – whereas acceptance of H (or even a large p-value) does not provide any evidence in favor of the canonical model. Although many authors (see e.g. MacKinnon (1992) or Dette & Munk (1998a,b), Munk & Dette (1999), Munk & Czado (1998), Czado & Munk (1998) for the assesment of distributional assumptions and Hauck & Anderson (1997) for more general models) have criticized to treat the problem of model selection as a testing problem of a point null hypothesis H such as in (1), this way of proceeding is still common practice among applied statisticians since alternative procedures are usually not developed.

Therefore, the main concern of this paper is to suggest the consideration of p-value surfaces and associated tests for precise hypotheses as alternatives. We will show that this gives more accurate information about the deviation from the canonical model.

To fix ideas, we restrict for the moment our consideration to the large class of generalized logistic links introduced by Czado (1992), even though any other parametric class mentioned above could have been used. However, for the analysis of the following two examples this class of Box-Cox transformations of the linear predictor is preferred because it allows separate modification of the right and left tail of the link function and its parametrization is locally orthogonal (Czado (1997)). In particular, the family allowing for a right tail modification $\mathcal{F}_R = \{F(\cdot,\psi), \psi \in \Re\}$ is given by:

$$F(\eta,\psi) = \frac{\exp(h(\eta,\psi))}{1+\exp(h(\eta,\psi))}, \quad \text{where } h(\eta,\psi) = \begin{cases} \frac{(\eta+1)^{\psi}-1}{\psi} & \text{if } \eta > 0 \\ \eta & \text{otherwise} \end{cases} . \tag{2}$$

A family allowing for a left tail modification can be defined similarly. Note, that for this family $\psi = \psi_* = 1$ always corresponds to the canonical logistic link. For $\psi < 1$ ($\psi > 1$) the right tail is heavier (lighter) than the logistic distribution ($\psi = 1$). These links have low variance inflation (Taylor (1988)) due to the fact that the parametrization is orthogonal in a neighborhood around $\beta = 0$. In addition, they are location and scale invariant (cf. Czado (1997)).

Example 1: (Age of Menarche in Warsaw Girls) Milicer and Szczotka (1966) analyzed the occurrence of menarche as a function of age (see Table 3 of Stukel (1988) for data). The standard logistic analysis with age as covariate reveals lack of fit in the left tail. Table 1.1 gives parameter estimates and their estimated standard errors in parentheses in the first column. Residual deviances, their degrees of freedom and the p-value of corresponding goodness of fit test in parentheses are given in the second column. The likelihood ratio (LR) statistic for testing logistic link by (1), their degree of freedom and the corresponding p-value in parentheses is provided in the third column. Therefore, following the usual way of proceeding we would decide for a noncanonical model with left tail modification.

Model	Estimated Link ψ	Residual Deviance	Likelihood Ratio
logistic		26.70 (23, .27)	
right tail	.88 (.083)	25.09 (22, .29)	1.61 (1, .204)
left tail	1.39 (.138)	17.62 (22, .73)	9.08 (1, .003)

Table 1.1: Link Estimates, Residual Deviance and LR Statistics for the Age of Menarche Data

Example 2: (Bottle Deposit Data) Neter, Wasserman & Kutner, p. 617 (1989) gave data on the number of bottles returned for 6 different levels of deposits. The results of a generalized logistic analysis are contained in Table 1.2.

logistic		12.18 (4, .02)	
right tail	1.56 (.231)	5.28 (3, .15)	6.90 (1, .009)
left tail	.63 (.197)	10.03 (3, .02)	2.15 (1, .143)

Table 1.2: Link Estimates, Residual Deviance and LR Statistics for the Bottle Deposit Data

Here we are left in a somewhat difficult situation. Although the LR-test supports a right tail modification with high evidence (p-value = .009) the residual deviance only indicates a slight gain in fit (p-value = .15). The LR-test could be too sensitive (detecting small departures from the canonical model which are scientifically irrelevant) or the residual deviance test could be not powerful enough.

In order to overcome those ambiguous situations arising from testing H in (1), we suggest in a first step to determine a measure of discrepancy Δ between the canonical and link in terms of quantities which are *scientifically relevant* for the experimenter. For example, the cost of a noncanonical link (as described above) need not to be justified, if the effects of using this link instead of the more "fitting" noncanonical link are small on the mean response estimates or other quantities of particular interest. We will see in Section 4, that the effects on the mean response estimates are about the same for both data sets, although the effects are very different on the estimated odds. More specifically, the odds are changed up to a factor of 50 for the menarche data set, while they are changed only up to a factor of 1.5 for the deposit data. Therefore, if the odds is the parameter of interest, a noncanonical link is truly needed for the menarche data, while it is not necessary for the deposit data. One could also be interested in other quantities as the odds which affects the above conclusions. The choice of such alternative measures of discrepancy between the canonical and noncanonical model will be discussed carefully in Section 3.

In a second step we suggest to consider generalized p-value curves associated to one sided tests for Δ. For a broad class of discrepancy measures Δ it is shown, that this is tantamount to test precise hypotheses of the form

$$H : \psi \notin [\psi_* - \psi_l, \psi_* + \psi_u] \qquad \text{versus} \qquad K : \psi \in [\psi_* - \psi_l, \psi_* + \psi_u] \tag{3}$$

for specified $0 < \psi_l, \psi_u$. If H is rejected at some level α, the canonical link is validated with controlled error probability α within a $(\psi_* - \psi_l, \psi_* + \psi_u)$-neighborhood. In Section 3, we show how these bounds can be derived from Δ. In particular, the evidence of H and K can simply be graphically illustrated by these p-value curves. Simple quantities of these curves, such as steepness, allow to visualize rapidly the goodness of fit – for both, validation and discrimination. We indicate in Section 5 how these curves can have a valid Bayesian interpretation as measures of evidence for H and K.

Our approach is based on the asymptotic distribution of the joint maximum likelihood estimator of link and regression parameters. For this we extend in Section 2 results by Fahrmeir and Kaufmann (1985) and apply these to the construction of tests and p-value curves for the problem (3). The simulation results in Appendix A of this test for validating and discriminating a logistic link show that the asymptotic law is a quite good approximation in small samples, which allow the proposed tests to be used for the analysis of a link in GLM at controlled error rate. In Section 4, we return to the examples presented above and illustrate generalized p-value curves in action. Finally, it should be noted, that a link misspecification represents only a special systematic departure from the model, while misspecification of the variance function or scales of the covariates are other possible departures. In this paper we focus mainly on link misspecifications, however, in Section 5 it is indicated for the negative binomial regression model for overdispersion (cf. Lawless, 1987) how the proposed methodology can be transferred to the assessment of other departures from the model.

Asymptotic Theory. Ordinary GLM's have been extended to allow for data selected link functions from a class of parametric functions. For binomial responses, this is evidenced by the many parametric link families considered in the literature. In the context of other GLM's, this extension was first considered by Pregibon (1980) and investigated in more detail by Czado (1992, 1997). The following model for regression data with response Y_i and explanatory variables $\boldsymbol{X}_i = (x_{i1}, \cdots x_{ip})$ for $i = 1, \cdots, n$ will be used:

1. **Random Component:** $\{Y_i, 1 \leq i \leq n\}$ are independent and have density of the form
$$f_{y_i}(y_i, \theta_i, \phi) = \exp[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)] \tag{4}$$
for some specified functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. The scale parameter ϕ is allowed to be known or unknown.

2. **Systematic Component:** The linear predictors $\eta_i(\beta) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ for $1 \leq i \leq n$ influence the response Y_i. Here $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$ are unknown regression parameters.

3. **Parametric Link Component:** The linear predictors $\eta_i(\beta)$ are related to the mean μ_i of Y_i by $\mu_i = F(\eta_i(\beta), \psi)$ for some $F(\cdot, \psi)$ in $\Im = \{F(\cdot, \psi) : \psi \in \Psi\}$.

Attention is restricted to link families $\Im$ which contain only strictly monotone continuous functions $F(\cdot, \psi)$ indexed by a scalar link parameter ψ. Note that in conventional GLM terminology the link g is equal to the inverse of F. An unknown scale parameter ϕ in (4) is estimated by an appropriate moment estimator involving the Pearson χ^2 Statistic. In GLM's with parametric link (see 4), the regression parameter $\boldsymbol{\beta}$ and the link parameter ψ are jointly estimated by maximum likelihood. If the true link F is a member of the link family $\Im$, the joint MLE $\hat{\boldsymbol{\delta}} = (\hat{\beta}, \hat{\psi})$ of $\boldsymbol{\delta} = (\beta, \psi)$ will be shown to be strongly consistent and efficient under regularity conditions. This asymptotic normal distribution of the joint MLE $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}, \hat{\psi})$ of $\boldsymbol{\delta} = (\beta, \psi)$ can then be used to construct a validation test for H versus K in (3).

As for ordinary GLM's, one has the relationship $\mu_i = \frac{d}{d\theta} b(\theta)|_{\theta=\theta_i} = b'(\theta_i)$. The log likelihood $l(\delta)$ derived from model (4) can be written as:

$$l(\delta) = \sum_{i=1}^{n} [\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)] \text{ where } \mu_i = b'(\theta_i) \text{ and } \mu_i = F(\eta_i(\beta), \psi).$$

To derive the corresponding scores, note that $\mu_i = F(\eta_i(\beta), \psi)$ holds, which implies

$$\frac{\partial \mu_i}{\partial \psi} = F_{i2}\,, \frac{\partial \mu_i}{\partial \beta_j} = x_{ij} F_{i1} \quad \text{for } \; 0 \leq j \leq p, 1 \leq i \leq n, \tag{5}$$

where $F_{i1} = \frac{\partial}{\partial \eta} F(\eta, \psi)|_{\eta=\eta_i}, F_{i2} = \frac{\partial}{\partial \psi} F(\eta, \psi)|_{\eta=\eta_i}$ and $x_{i0} = 1$ for $1 \leq i \leq n$. Let $d_i = \frac{d\theta_i}{d\mu_i}$ and use (5) to express the scores as follows:

$$\begin{aligned} s_j(\boldsymbol{\delta}) &= \frac{\partial}{\partial \beta_j} l(\delta) = \sum_{i=1}^{n} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \beta_j} [\frac{y_i - \mu_i}{a(\phi)}] = a(\phi)^{-1} \sum_{i=1}^{n} d_i x_{ij} F_{i1}(y_i - \mu_i), \qquad (6) \\ s_{p+1}(\boldsymbol{\delta}) &= \frac{\partial}{\partial \psi} l(\delta) = \sum_{i=1}^{n} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \psi} [\frac{y_i - \mu_i}{a(\phi)}] = a(\phi)^{-1} \sum_{i=1}^{n} d_i F_{i2}(y_i - \mu_i). \end{aligned}$$

zero. Finally, the expected Fisher information $I_n(\delta)$ for model (4) can be expressed as follows:

$$I_n(\boldsymbol{\delta}) = a(\phi)^{-1} \left[\begin{array}{cc} I_{\beta,\beta} & I'_{\beta,\psi} \\ I_{\beta,\psi} & I_{\psi,\psi} \end{array} \right], \tag{7}$$

where $I_{\beta,\beta}$ is a (p+1)x(p+1) matrix, $I_{\beta,\psi}$ is a (p+1) vector and $I_{\psi,\psi}$ is a real number given by

$$(I_{\beta,\beta})_{rs} = \sum_{i=1}^{n} x_{is} x_{ir} F_{i1}^2 d_i, \quad (I_{\beta,\psi})_r = \sum_{i=1}^{n} x_{ir} F_{i1} F_{i2} d_i \;\; 0 \leq r,s \leq p \quad \text{and } I_{\psi,\psi} = \sum_{i=1}^{n} F_{i2}^2 d_i.$$

It is easy to see, that the score vector $s_n(\boldsymbol{\delta}) = (s_1(\boldsymbol{\delta}), \cdots, s_{p+1}(\boldsymbol{\delta}))$ has covariance matrix $I_n(\boldsymbol{\delta})$. Let $H_n(\boldsymbol{\delta})$ denote the corresponding observed information matrix with (s,t)th element given by

$$H_n(\boldsymbol{\delta})_{st} = \frac{\partial^2 l(\boldsymbol{\delta})}{\partial \delta_s \delta_t} \text{ for } s,t = 1, \cdots, p+1.$$

The minimal (maximal) eigenvalue of a square matrix A will be written as $\lambda_{min}(A)(\lambda_{max}(A))$. Let $\delta_0 = (\beta_0, \psi_0)$ denote the true parameter values. For brevity, we will write I_n and s_n for $I_n(\delta_0)$ and $s_n(\delta_0)$, respectively. The following regularity conditions are needed:

R1 $\lambda_{min}(I_n) \to \infty$ as $n \to \infty$.

R2 There is a neighborhood $N \subset B$ of $\boldsymbol{\delta}_0$ such that a.s

$$\lambda_{min}(H_n(\delta)) \geq c[\lambda_{max} I_n]^{\frac{1}{2}+\varepsilon}, \delta \in N, n \geq n_1$$

with some constants $c, \varepsilon > 0$ and a random number n_1.

R3 Assume $\{\mathbf{x}_n, n \geq 1\} \subset K$ compact and $F(x_n^t \beta, \psi)$ twice differentiable with respect to $\boldsymbol{\beta}$ and ψ and bounded for $\{\mathbf{x}_n, n \geq 1\} \subset K$ for fixed β and ψ,

R4 $\frac{n}{\lambda_{min}(I_n)}$ is uniformly bounded $\forall n \geq 1$.

The following results are modifications of results for ordinary links previously obtained by Fahrmeir & Kaufmann (1985). Observe, that additional estimation of the link requires slightly stronger assumptions on the link function $F(\cdot, \psi)$. Here $F(\eta, \psi)$ has to be twice differentiable with regard to ψ.

Theorem 2.1 *Under (R1) and (R2) with* $\varepsilon > 0$, *there is a sequence* $\hat{\boldsymbol{\delta}}_n$ *of random variables and a random number* n_1 *with*

(i) $P(s_n(\hat{\boldsymbol{\delta}}_n) = \mathbf{0}$ *for all* $n \geq n_1) = 1$ *(asymptotic existence),*

(ii) $\hat{\boldsymbol{\delta}}_n \to \boldsymbol{\delta}_0$ *a.s. (strong consistency)*

We are now in the position to give the asymptotic result:

Theorem 2.2 *Under (R1), (R3) and (R4), there is a sequence* $\hat{\boldsymbol{\delta}}_n$ *such that* $P(s_n(\hat{\boldsymbol{\delta}}_n) = \mathbf{0}) \to 1$ *as* $n \to \infty$ *and*

$$I_n^{\frac{1}{2}}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{\mathcal{D}} N_{p+1}(\mathbf{0}, I) \qquad \text{as } n \to \infty,$$

where $N_m(\mu, \Sigma)$ *denotes an m-dimensional normal distribution with mean vector* μ *and variance-covariance matrix* Σ.

Asymptotic Link Validation Tests in Generalized Linear Models. Using the above results, we are now able to construct an asymptotic link validation test for GLM's.

Theorem 2.3 *Under the assumptions of Theorem 2.2 for the validation problem H versus K in (3) a consistent asymptotic level α test is given by the rejection region*

$$\mathcal{C} \;=\; \left\{\hat{\psi}_n \;:\; \left|\Phi\left(\frac{\hat{\psi}_n - \psi_* - \psi_u}{\hat{\sigma}_n(\psi,\beta)}\right) \;-\; \Phi\left(\frac{-(\hat{\psi}_n - \psi_*) - \psi_l}{\hat{\sigma}_n(\psi,\beta)}\right)\right| \le \alpha\right\} \tag{8}$$

where $\hat{\sigma}_n^2$ denotes any consistent estimate of $\sigma_n^2(\psi,\boldsymbol{\beta})$ which is the $(p+1,p+1)$-th element of the inverse of the Fisher information matrix $I_n(\boldsymbol{\delta})$. The critical region of a test at level $1-\alpha$ for the discrimination problem K versus H is given by the complement $\mathcal{C}^c$.

Proof.

Fix $\hat{\sigma}^2$. We first symmetrize the problem. For this define

$$\theta = \psi - \psi_* - \frac{\psi_u - \psi_l}{2} \text{ and } \theta_1 = \frac{\psi_u + \psi_l}{2}.$$

Therefore H in (3) is equivalent to $H : \theta \in [-\theta_1, \theta_1]$. Since $\hat{\theta}_n = \hat{\psi}_n - \psi_* - \frac{\psi_u - \psi_l}{2}$ has an approximate normal distribution, we construct as in Lehmann (1986), Th.6, p.101 an equivalence test for the symmetrized problem. The critical region of this test is given by

$$\mathcal{C} := \left\{\hat{\theta}_n \;:\; \hat{\theta}_n \in [-C, C]\right\}, \text{ where } C > 0 \text{ is uniquely determined by}$$

$$\hat{P}(C, \hat{\sigma}_n, \psi_l, \psi_u) := \Phi\left(\frac{C - \theta_1}{\hat{\sigma}_n}\right) \;-\; \Phi\left(\frac{-(C + \theta_1)}{\hat{\sigma}_n}\right) = \alpha. \tag{9}$$

Note further, that condition (25) of Lehmann (1986, p.102) reduces to (9) out of symmetry. Since

$$\Phi\left(\frac{\theta - \theta_1}{\hat{\sigma}_n}\right) \;-\; \Phi\left(\frac{-(\theta + \theta_1)}{\hat{\sigma}_n}\right)$$

is monotone increasing in θ and zero for $\theta = 0$, condition (9) is equivalent to

$$\mathcal{C} \;=\; \left\{\hat{\theta}_n \;:\; \left|\Phi\left(\frac{\hat{\theta}_n - \theta_1}{\hat{\sigma}_n}\right) \;-\; \Phi\left(\frac{-(\hat{\theta}_n - \theta_1)}{\hat{\sigma}_n}\right)\right| \le \alpha\right\}.$$

This last statement can be rewritten as (8). From Theorem 2.2 we conclude that

$$\mathcal{L}\left\{(\hat{\psi}_n - \psi)\sigma_n^{-1}(\beta, \psi)\right\} \;\longrightarrow\; \mathcal{N}(0,1) \qquad \text{as } n \to \infty.$$

Applying Slutzky's Theorem proves that the test is asymptotic size α. Consistency is similar.

To apply the last theorem we have to estimate $\sigma_n^2(\psi, \beta)$ by $\hat{\sigma}_n^2 := \sigma_n^2(\hat{\psi}, \hat{\beta})$, where $(\hat{\psi}, \hat{\beta})$ is the joint MLE of (ψ, β).

Tolerance Bounds. Crucial for the specification of the hypotheses H and K in (3) are the values of the tolerance constants $\psi_l > 0$ and $\psi_u > 0$. We will now discuss several choices depending on the quantities one is interested in estimating and the assumed GLM model. In order to illustrate a general strategy let $\hat{\beta}(\psi)$ denote the MLE of β, when a fixed link parameter ψ of an arbitrary parametric link family $\mathcal{F} = \{F(\cdot, \psi), \psi \in \Psi\}$ is used and $\eta_i(\hat{\beta}(\psi))$ denotes the corresponding i-th linear predictor.

Step 1. *Determine a measure of discrepancy $\Delta_{(\cdot)}$ from the canonical model.*

This can be realized e.g. for all GLM's by the maximal change in the mean response estimates when switching from the canonical link ψ_* to the link ψ, i.e. $\Delta_{(\cdot)} = \Delta_m$, where

$$\Delta_m(\psi) \;\; = \;\; \max_{i=1,\cdots,n} |F(\eta_i(\beta(\psi)), \psi) - F(\eta_i(\beta(\psi_*)), \psi_*)|. \tag{10}$$

Here $\beta(\psi)$ denotes the true (unknown) regression parameter in the model $F(\cdot, \psi)$. Since $\beta(\psi)$ are unknown, we will use $\hat{\beta}(\psi)$ the MLE of β for fixed ψ as an estimate for $\beta(\psi)$. As noted by a referree, the discrepancy between two members $F(\cdot, \psi_1)$ and $F(\cdot, \psi_2)$ can also be considered. It is straight forward to extend the dicrepancy measure from the canonical model to study the last question.

Step 2: *Determine the corresponding tolerance bound Δ_0.*

For this bound Δ_0 the canonical model is assumed as sufficiently approximated by the noncanonical model. Testing $\tilde{H}: \; \Delta_m(\psi) > \Delta_0$ (or $\tilde{K}: \; \Delta_m(\psi) \leq \Delta_0$) is now tantamount to the testing problem (3) and its converse, where the bounds $\psi_* - \psi_l, \psi_* + \psi_u$ have to be determined numerically as $\Delta_m^{-1}(\Delta_0)$. Observe, that this leads always to two unique values $\psi_* - \psi_l^0 < \psi_* + \psi_u^0$, such that $\psi_* - \psi_l^0 < \psi < \psi_* + \psi_u^0$ because the criterion $\Delta_m(\psi)$ is strictly unimodal with unique minimum at ψ_*. ψ_l, ψ_u can now be expressed in terms of Δ_m in accordance with the particular model and the specific question the experimenter has in mind.

This proceeding applies, of course, to other measures of discrepancy. For example, often not only absolute changes in the mean responses are of interest, but also relative changes

$$\Delta_r(\psi) = \max_{i=1,\cdots,n} \begin{cases} \frac{F(\eta_i(\beta(\psi)),\psi)}{F(\eta_i(\beta(\psi_*)),\psi_*)} & \text{if } F(\eta_i(\beta(\psi)), \psi) < F(\eta_i(\beta(\psi_*)), \psi_*) \\ \frac{F(\eta_i(\beta(\psi_*)),\psi_*)}{F(\eta_i(\beta(\psi)),\psi)} & \text{otherwise} \end{cases} . \tag{11}$$

In order to guarantee that assessment of the model with respect to such a criterion $\Delta_{(\cdot)}$ can be treated by two sided hypotheses as in (3) it is sufficient that $\Delta_{(\cdot)}(\psi)$ is a strictly unimodal function with unique minimum at ψ_*. Note, that this property holds for the above criteria as well as for the following ones. We discuss now some criteria which are more specifically adapted to particular GLM's.

Binomial Responses. For binomial responses covariate effects are often interpreted using odds ratio's. Therefore we consider the maximal change in the odds which can be estimated by:

$$\Delta_o(\psi) = \max_{i=1,\cdots,n} \begin{cases} o_i(\psi) & \text{if } F(\eta_i(\beta(\psi)), \psi) > F(\eta_i(\beta(\psi_*)), \psi_*) \\ \frac{1}{o_i(\psi)} & \text{otherwise} \end{cases} , \tag{12}$$

where

$$o_i(\psi) = \left[\frac{F(\eta_i(\beta(\psi)), \psi)}{1 - F(\eta_i(\beta(\psi)), \psi)}\right] / \left[\frac{F(\eta_i(\beta(\psi_*)), \psi_*)}{1 - F(\eta_i(\beta(\psi_*)), \psi_*)}\right].$$

allows to weigh differences for tail success probabilities heavier than success probabilities around .5.

As a lower bound for (10) we can compare the maximal absolute difference between all possible success probabilities under a logistic model and a model with link parameter ψ, given by

$$\Delta_p(\psi) = sup_{\eta \in \mathcal{R}} |F(\eta, \psi) - F(\eta, \psi_*)|. \tag{13}$$

Since $F(\eta, \psi)$ is bounded the supremum in (13) is finite. In order to illustrate this measure Figure 3.1 gives the absolute difference in probability between the logistic and the generalized logistic distribution as a function of η and ψ. It can be seen that for $\psi > 1$ (lighter right tail) this difference is significantly large in a much smaller range of η values compared to the case of $\psi < 1$ (heavier right tail). This allows us to classify four areas of varying degree of information about ψ. In the case of a heavier right tail ($\psi < 1$) compared to the logistic link ($\psi = 1$) *and* a large range for the linear predictors η_i it will be easy to discriminate against the logistic link, while the opposite will be true in the case when there is small range for the linear predictors. For the lighter right tail case ($\psi > 1$), the degree of information for discriminating against the logistic link will be medium in both cases of a large or small range for the linear predictors η_i. The maximal distance $\Delta_p(\psi)$ between the generalized logistic link and the logistic link as a function of ψ is given in Figure 3.2. Table 3.1 gives for some special Δ_0 values the corresponding interval $[\psi_* - \psi_l, \psi_* + \psi_u]$ to insure that $\Delta_p(\psi) < \Delta_0$ for all $\psi \in [\psi_* - \psi_l, \psi_* + \psi_u]$.

	$\Delta_p(\psi) = \Delta_0$						
	.01	.025	.05	.075	.1	.15	.20
$\psi_* - \psi_l$	.93	.82	.65	.49	.35	.02	-.28
$\psi_* + \psi_u$	1.07	1.19	1.39	1.60	1.84	2.40	3.10

Table 3.1: Choice of lower and upper bounds for ψ to achieve a maximal absolute difference of Δ_0 between the generalized logistic cdf and logistic cdf

Poisson Responses. In this case one might be interested in determining the change in probabilities of no event occurring given by

$$\Delta_n(\psi) \quad = \max_{i=1,\cdots,n} |\exp(-F(\eta_i(\beta(\psi)), \psi)) - \exp(-F(\eta_i(\beta(\psi_*)), \psi_*))| \tag{14}$$

with lower bound

$$\Delta_{np}(\psi) = sup_{\eta \in \mathcal{R}} |\exp(-F(\eta, \psi)) - \exp(-F(\eta, \psi_*))|. \tag{15}$$

Note, that this leads to a similar surface as in Figure 3.1. As for binomial responses the changes in the odds of the probabilities of no event can also be considered

$$\Delta_{no}(\psi) = \max_{i=1,\cdots,n} \begin{cases} o_i^n(\psi) & \text{if } F(\eta_i(\beta(\psi)), \psi) < F(\eta_i(\beta(\psi_*)), \psi_*) \\ \frac{1}{o_i^n(\psi)} & \text{otherwise} \end{cases}, \tag{16}$$

where

$$o_i^n(\psi) = \left[\frac{\exp(-F(\eta_i(\beta(\psi)), \psi))}{1 - exp(-F(\eta_i(\beta(\psi)), \psi))}\right] / \left[\frac{\exp(-F(\eta_i(\beta(\psi_*)), \psi_*))}{1 - exp(-F(\eta_i(\beta(\psi_*)), \psi_*))}\right].$$

***p*-value Curves and Surfaces**. In addition to a pure test decision for a fixed bound Δ_0 we suggest consideration of the function $\hat{P}$ in (9). We will now show how this function can be utilized for the

the validation and discrimination problem (H versus K and converse) in (3). Observe, that given a fixed sample of observations $\hat{P}(\hat{\psi}_n, \hat{\sigma}_n, \psi_l, \psi_u)$ is a two – dimensional surface, where the level sets $\alpha = \hat{P}$ give the asymptotic minimal bounds $\psi_* - \psi_l$ and $\psi_* + \psi_u$ for which H can be rejected at level α as well as the maximal bounds for which K can be rejected by the discrimination test at level $1 - \alpha$. In particular, when $\psi_* - \psi_l = \psi_u = \psi_* + \psi_*$, 1-$\hat{P}$ denotes the 'classical' p-value of the maximum likelihood test for K against H. We suggest to consider $\hat{P}$ and its complement as an (asymptotically) *precise* measure of the evidence of neighborhoods $(\psi_* - \psi_l, \psi_* + \psi_u)$ in contrast to the classical 'two-sided' p-value associated to (1). For an illustration of $\hat{P}$ we defer to the examples discussed in Section 4.
As we will see in Section 5 it is crucial for a valid interpretation of the proposed p-value approach, that these two-dimensional surfaces $\hat{P}$ reduces to one-dimensional curves with respect to the particular discrepancy measure $\Delta_{(\cdot)}$. Once decided for a criterion as $\Delta_m(\psi), \Delta_o(\psi)$ or $\Delta_p(\psi)$, this surface only depends on $\Delta_{(\cdot)}$ by the relation $\hat{P}(\cdot, \cdot, \Delta_{(\cdot)}) := \hat{P}(\cdot, \cdot, \psi_l, \psi_u)$. For illustration in the case of generalized logistic regression for the criterion Δ_p confer Table 3.1 again.

4 Examples

Binomial Responses

Age of Menarche in Warsaw Girls (Revisited). For this data set, changes to the estimated success probabilities (see (13)) as well as to the estimated odds (see (12)) have been investigated. First, for a range of Δ_0 values the corresponding $\psi_* - \psi_l$ and $\psi_* + \psi_u$ values have been determined for both criteria. The corresponding generalized p-value functions (as defined in Section 2.2) are given in Figure 4.1. These functions were calculated as functions of the particular criterion $\Delta_{(\cdot)}$.

They show, that using a left tail link modification will result in a maximal absolute difference of 3% in estimated success probabilities at $\alpha = .1$ compared to a logistic analysis. A logistic analysis can be validated at $\alpha = .1$, if one is willing to tolerate a change of 7% in estimated probabilities. Since this data set contains extreme observed probabilities, it will be expected that the effects on the estimated maximal odds will be large, which is supported by Figure 4.1. In particular, a maximal change of the estimated odds by 5 can be detected, but the logistic link can only be validated when accepting maximal change by 50. Given these results, it seems to be reasonable that a noncanonical link model is necessary if interest is focused on the odds. This is in accordance to the analysis made by the standard tests in the introduction. However, if we are only interested in the maximal probability difference, a modification of the model seems to be unnecessary, because a maximal probability difference under a left tail modification of ≤ 0.07 and ≤ 0.03 using a right tail modification can be validated at .1. Note, that consideration of the p-value for a test of (1) does not allow such a conclusion.

Bottle Deposit Data (Revisited) Remember, that the standard LR test gives strong indication for a noncanonical link model. As for the menarche data changes to the estimated mean responses as well as to the estimated odds are considered and the results are plotted in Figure 4.2. It shows, that a right tail modification will result in a maximal difference of 3% in estimated success probabilities compared to a logistic link analysis. This analysis can be validated at $\alpha = .1$, if one tolerates a change of 10% in probabilities. The maximal change on estimated odds is much less compared to the menarche data set. Here the logistic link can be validated in the neighborhood of a maximal change on the estimated odds of 1.48 at $\alpha = .1$. If the emphasis is on estimating odds, this change is certainly tolerable. Therefore, if the parameter of interest is the odds, it is certainly justified to maintain a logistic link despite the observed significant improvement in fit by the classical goodness of fit statistics, when a right tail modification and the estimated variability in the link is small.

Poisson Responses

Mining Fracture Data. Myers (1990, p. 336) reports on the number of injuries or fractures that occur in the upper seam of mines in the coal fields of the Appalachian region in western Virginia. Four potential covariates were collected. Myers suggests that a Poisson regression with canonical (log) link utilizing three covariates (inner burden thickness, the percentage of extraction of lower previously mined seam and the time in years that the mine has been in operation) provides a reasonable model. Since the observed counts are positive it is more important to validate the canonical link against a right tail modification than a left tail one. For this data set, the residual deviance for the canonical link ($\psi = 1$) is 38.03 (df=40), while for a right tail modification with estimated link parameter $\hat{\psi} = -.64(.13)$ the residual deviance is 28.21 (df=39), indicating that a link modification might be necessary. This is confirmed by Figure 4.5.

This shows that using a right tail link modification will result in a maximal absolute difference of 1.05 in estimated mean responses at $\alpha = .1$ compared to a canonical link analysis. This change is very large since the median number of injuries reported is 2. The maximal possible change in no event probabilities is .115 and the maximal change for estimated odds of probabilities of no injuries is 280 at $\alpha = .1$. Therefore the canonical link cannot be validated and a right tail modification is truely needed.

5 Discussion and further remarks

p-value Curves as a Measure of Evidence. That p-values for simple null hypotheses as well as for precise hypotheses cannot be considered as a measure of evidence (in the sense of Bayesian posterior probabilities) has been forcefully shown by many authors (see e.g. Berger & Delampady (1987), Delampady, (1989)). This is supported for the particular problem of choosing a link function in the simulation study in Appendix A. Certainly, testing precise hypotheses as in (3) does not solve the problem of the assessment of the correct model in the above sense. However, these hypotheses allow at least for a much more flexible possibility to guard against misspecifications towards a *direction* which is considered as the more serious error (against the canonical model or in favor of the canonical model). Moreover, this direction can be expressed and quantified in terms of entities (such as the odds) which are of primary interest for the experimenter. In contrast, the classical null (1) forces us, to consider the rejection of the canonical model although being true, as the more serious error. In many cases the type II error seems, however, to be the more serious one, because acceptance of the canonical model will heavily affect the subsequent data analysis (for example the MLE will be different). Here, at least for frequentists, the formulation (3) is more appropriate.

From our own experience we know that in many applications it is not obvious what the type I and type II error should be – therefore, an explorative data analysis may become more attractive. In some sense, p-value curves, as suggested in Section 3, may be regarded as a sort of standardized EDA analysis.

Nevertheless, practitioners often like to attach the meaning 'the probability that the null H is true' to p-values (Casella & Berger (1987)). Indeed, the following argument shows that for the criteria suggested in Section 3 the associated p-value curves provide additional knowledge about the *evidence* of the canonical model. For this observe, that the choice of one of the measure of discrepancy $\Delta_{(.)}$ in Section 3 always implies that the *two-sided* test problem (3) (or its converse) reduces to a *one-sided* test problem concerning the parameter $\Delta_{(.)}(\psi)$. Hence, we are (at least approximately) in the situation of a one sided test problem for the unknown location parameter $\Delta_{(.)}$ in a normal model. Now, p-value curves $\hat{P}(\Delta_{(.)})$

(1987)) for the probability that the canonical model is true given some prior $\pi(\Delta_{(.)})$. Schervish (1996) showed further that in the normal case, p-value curves can be interpreted as penalizing hypotheses that contain additional parameters (in the sense of Schervish's Definition 1) that are far away from the data. Hence, *p-value curves* for $\Delta_{(.)}$ represent a useful guide for the task of practitioners to interpret the outcome of a pivot statistics as a measure of evidence.

Effects of Isolated Departures from the Model. Data on the effects of insecticides on flour beetles presented in Collett (1991, p. 142) provide an example that care has to be taken in the presence of isolated departures from the canonical model. Fitting parallel lines for the insecticide effects, the results of a right and left tail modification of the logistic link for the complete data and the data with the outlier removed are presented in Table 5.1. While there is no evidence for link misspecification, it shows that the residual deviance is inflated when the outlier is present. Here, the residual deviance is oversensitive to isolated departures, while the LR statistic is not.

p-value curves for the probabilities (Figure 5.1) and odds (not shown) show that the logistic link can be validated in a neighborhood of 6 % (10.5%) maximal change to the success probabilities for the complete data (data with outlier removed). This relatively large neighborhood is the result of the medium size of n_i. While the LR statistic gives the impression of a perfect fit to the logistic link, the asymptotic link validation test shows some uncertainty due to sparse information about the link.

	Complete Data			Outlier removed		
Model	Link Estimate	Residual Deviance	Likelihood Ratio	Estimate Estimate	Residual Deviance	Likelihood Ratio
logistic		21.28 (14, .08)			14.84 (13, .32)	
right tail	1.20 (.25)	21.45 (13, .06)	.68 (1, .41)	1.12 (.23)	14.55 (12, .27)	.29 (1, .59)
left tail	1.43 (.37)	21.05 (13, .07)	1.1 (1, .30)	1.14 (.30)	14.65 (12, .26)	.19 (1, .66)

Table 5.1: Link Estimates, Residual Deviances and LR Statistics for the Flour Beetle Data

This shows that isolated departures do certainly influence the performance of the link validation test. It is to be expected that missing covariates and overdispersion in the data also influence the validation test. Therefore the link validation test should only be applied after diagnostic tools for the detection of a mean misspecification such as developed by Landwehr, Pregibon and Shoemaker (1984), Pregibon (1981), Williams (1987) and O'Hara, Hines and Carter (1993) have been used. In the presence of overdispersion, score point hypothesis tests developed by Dean (1992) and Smith and Heitjan (1993) can be applied. Appropriate interval hypothesis tests, however, would be preferable over these score tests by the same reasons as for testing the goodness of link. This we will briefly sketched in the following.

Overdispersion. We develop a validation test along the lines in Section 2 and 3 for the overdispersion parameter $a > 0$ in the special model of a negative binomial (NB) Poisson regression (cf. Lawless 1987). Let $\hat{\beta}(a)$ denote the MLE of the regression parameter β with fixed overdispersion parameter a, then the asymptotic covariance matrix for $\hat{\beta}(a)$ is given by $(X^t D(\mu(\beta(a)), a)X)^{-1}$, a diagonal matrix with ith element given by

$$D_i(\mu(\beta(a)), a) = \frac{\mu_i(\beta(a))}{1 + a\mu_i(\beta(a))}, \text{ where } \mu_i(\beta(a)) = T_i exp(x_i^T \beta(a)).$$

Here, $a \geq 0$ where $a = 0$ corresponds to no overdispersion. We will se in the following that it is reasonable to measure the effects of overdispersion by means of the asymptotic covariance matrix of $\hat{\beta}(a)$, which

$$\Delta_{over}(a) = \inf_{\mu \in \Re^n} \frac{||(X^t D(\mu, a)X)||}{||(X^t D(\mu, 0)X)||},$$

where $||A||$ denotes the determinant of a square matrix A and $D(\mu, a)$ is a diagonal matrix with ith entry given by $\frac{\mu_i}{1+a\mu_i}$.

Lemma 5.1 Δ_{over} *is a strictly decreasing function with maximum* $\Delta_{over}(0) = 1$ *and minimum* $\lim_{a\to\infty} \Delta_{over}(a) = 0$

(For a proof see Appendix B.)

Note that because of

$$\Delta_{over(a)} \leq \frac{||(X^t D(\mu(\beta(a), a)X)||}{||(X^t D(\mu(\beta(0), 0)X)||}$$

$\Delta_{over}(a)$ is a lower bound for the ratio of the asymptotic volume of the confidence ellipsoids for $\hat{\beta}(\cdot)$ under the presence of overdispersion a and no overdispersion and hence testing $H : \Delta_{over}(a) \leq \Delta_0$ versus $K : \Delta_{over}(a) > \Delta_0$ for some specified $0 < \Delta_0 < 1$ will provide significant evidence for the closeness to the GLM without overdispersion. By means of Lemma 5.1 the corresponding asymptotic test and its p-value curve can be developed by constructing an asymptotic test for $H : a > a_0 = \Delta_{over}^{-1}(\Delta_0)$ versus $K : a \leq a_0$. Such a test, however, is given analogously to Theorem 2.3 using the asymptotic distribution of $\hat{a}$, the MLE of a, given in Lawless (1987, p. 211).

— Aranda-Ordaz, F.J. (1981) On Two Families of Transformations to Additivity for Binary Response Data, *Biometrika*,**68**, 357-364.

— Atkinson, A.C. (1982) Regression diagnostics, transformations and constructed variables (with discussion). *J. Roy. Statist. Soc. Ser. B*,**44**, 1-36.

— Berger, J.O., Delampady,M. (1987). Testing precise hypotheses. *Statistical Science* **2**(3), 317-52.

— Casella, G., Berger, R.L. (1987). Reconcilling Bayesian and Frequentist evidence in the one-sided testing problem. *Journ. Americ. Statist. Assoc.* **82**, 106-111.

— Collett, D. (1991). *Modelling binary data*, Chapman and Hall, London.

— Copenhaver, T.W. and Mielke, P.W. (1977). Quantit Analysis: A Quantal Assay Refinement, *Biometrics*, **33**, 175-186.

— Cox, D.R. and Reid, N. (1987) Parameter Orthogonality and Approximate Conditional Inference. *J. Roy. Statist. Soc. (B)*, **49**, 1-39.

— Czado, C. and Santner, T.J. (1992). The effect of link misspecification on binary regression inference. *J. Statist. Plan. Inf.***33**, 213-231.

— Czado, C. (1992). On Link Selection in Generalized Linear Models. in *Advances in GLIM and Statistical Modelling, Lecture Notes in Statistics*, **78**, Springer Verlag, New York.

— Czado, C. (1997). On Selecting Parametric Link Transformation Families in Generalized Linear Models. *J. Statist. Plan. Inf.* **61**, 125-141.

— Czado, C. and Munk, A. (1998). Assessing the similarity of distributions - finit sample performance of the empirical Mallows distance. *J. Statist. Comput. Simul.*, **60**, 319-346.

— Davison, A.C. and Tsai, C.-L. (1992). Regression Model Diagnostics, *Int. Statist. Rev.* , **60**, 3, 337-353.

— Dean, C.B. (1992) Testing for overdispersion in Poisson and binomial regression models. *J. Amer. Stat. Assoc.*, **87**, 451-457.

— Delampady, M. (1989). Lower bounds for Bayes factors for interval null hypotheses. *Jour. Americ. Statist. Assoc.* **84**, 120-24.

— Dette, H., Munk, A. (1998a). Validation of linear regression models. *The Annals of Statistics***26**, 778-800.

— Dette, H., Munk, A. (1998b). A simple goodness of fit test for linear models under a random design assumption. *Annals of Inst. Stat. Math.***50**,253-275.

— Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13**, 342-368 (Correction: *Ann. Statist.* **14**, 1643).

— Hauck, W.W., Anderson S. (1996). Discussion of: Bioequivalence trials, intersection union tests and equivalence confidence sets, by Berger,R.L., Hsu, J.C.. *Statistical Science*, **11**, 283-319.

— Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models. *J. Amer. Stat. Assoc.* , **79**, 61-83.

— Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *Canad. J. Statist.*,**15**, 209-225.

— Lee, A.H. (1987). Diagnostic Displays for assessing Leverage and Influence in Generalized Linear Models. *Austral. J. Statist.*, **29**, 233-243.

—Lee, A.H. (1988). Assessing Partial Influence in Generalized Linear Models, *Biometrics*, **44**, 71-77.

— Lehmann, E.L. (1986). *Testing Statistical Hypotheses.* 2nd ed. New York, John Wiley.

—O'Hara Hines, R.J. and Carter, E.M. (1993). Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. *Appl. Statistics*, **42**, 3-20.

— Guerrero, V.M. and Johnson, R.A. (1982). Use of the Box-Cox Transformation with Binary Response

— McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Second Edition, Chapman and Hall, London.

— MacKinnon, J.G. (1992). Model specification tests and artificial regressions. *Journal of Economic Lit.* **30**, 102-46.

— Milicer, H. and Szczotka, F (1966). Age at Menarche in Warsaw Girls in 1965, *Human Biology*, **38**, 199-203.

— Morgan, B.J.T. (1983). Observations on Quantit Analysis, *Biometrics*, **39**, 879-886.

— Munk, A. and Czado C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Statist. Soc. B*, **60**, 223-241.

— Munk, A. and Dette, H.(1998). Nonparametric comparison of several regression functions: exact and asymptotic theory, *Ann. Statist.*, **26**, 2339-2368.

— Myers, M (1992). *Classical and Modern Regression Analysis with Applications*, PWS-Kent.

— Neter, J., Wasserman, W., Kutner, M.H. (1989). *Applied Linear Regression Models*, Second Edition, Irwin, Boston.

– Pratt, J.W. (1965). Bayesian interpretation of standard inference statements (with discussion). *Jour. Roy. Statist. Soc. Ser. B* **27**, 169-203.

— Pregibon, D. (1980). Goodness of link tests for generalized linear models. *J. Roy. Statist. Soc. Ser. C* **29**, 15-24.

— Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist*, **9**, 705-724.

— Pregibon, D. (1982) Score Tests in GLIM with Applications. In GLIM82: *Proceedings of the International Conference on Generalized Linear Models, Lecture Notes in Statistics*, **13**. New York, Springer Verlag.

— Prentice, R.L. (1976) A Generalization of the Probit and Logit Methods for Dose Response Curves, *Biometrics*, **32**, 761-768.

— Schervish, M.J. (1996). P-values: what they are and what they are not. *The Americ. Statistician* **50**, 203-206.

— Smith, P.J. and Heitjan, D.F. (1993). Testing and adjusting for departures from nominal dispersion in generalized linear models. *Appl. Statistics*, **. 42, 31-41**.

—Stukel,T. (1988) Generalized logistic models. *J. Amer. Statist. Assoc.*, **83**, 426-431.

—Taylor, J.M.G. (1988). The cost of generalized logistic regression. *J. Amer. Statist. Assoc.*, **83**, 1078-1083.

—Van Montford, M.A.J. and Otten, A. (1976). Quantal Response Analysis: Enlargement of the logistic Model with a Kurtosis Parameter, *Biometrische Zeitschrift*, **18**, 371-380.

—Whittmore, A.S. (1983). Transformations to Linearity in Binary Regression, *SIAM Journal on Applied Mathematics*, **43**, 703-710.

—Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statistics*, **36**, 181-191.

Appendix A (Simulation Results)

Binomial Responses

Small Sample Behavior of Goodness of Fit Point Hypothesis Tests in a Logistic Model. To investigate the small sample behavior of the residual deviance as goodness of fit statistics and the LR statistics for testing a logistic link, we utilize the family of generalized logistic distributions defined in

$$Y_i \simeq \text{binomial}(n, p_i) \text{ for } i = 1, \cdots, 11 \text{ with } p_i = F(\beta_0 + \beta_1 x_i, \psi),$$

where $F(\eta, \psi)$ is given by (2) and $\beta_0 = 1$. Two choices for $\boldsymbol{x} = (x_1, \cdots, x_{11})$ were considered, $\boldsymbol{x}$ equally spaced between -5 and 5 and $\boldsymbol{x}$ a standard normal random sample of size 11 and β_1 was set to .5,1, or 2. This allows models with nearly symmetric true probabilities around .5 ($\beta_1 = 1$), extreme probabilities ($\beta_1 = 2$) and more central probabilities ($\beta_1 = .5$). Note that the equally (unequally) spaced covariate case will induce a large (small) range for the linear predictors. Therefore, we classify the cases $\psi < 1$ and equally spaced covariates ($\psi < 1$ and unequally spaced covariates) as areas with high (low) power to discriminate against the logistic link. The other areas ($\psi > 1$ and both cases of covariate configuration) have medium discrimination power. This will be supported by the following simulation results. Finally, we investigated two binomial sample sizes of $n = 20$ and $n = 40$.

To demonstrate the inappropriateness of using a large p-value of the ordinary goodness of fit statistics as an indicator of a good fitting model, we simultaneously calculated the p-values of the LR test of testing $\psi = 1$ as well as the residual deviance test assuming a logistic model based on 500 replications. Values for ψ were chosen between .02 and 2.4 to allow up to 15% percent of absolute difference in the probabilities between the logistic and generalized logistic model. We recorded the percentage of cases, where the p-value of the LR test statistic was larger than .2 (see Figure A.1) and .5 (not shown), respectively and the percentage where the p-value of the residual deviance statistic was larger than .5 (see Figure A.2) and .75 (not shown), respectively based on 500 replications. Different p-value bounds for the LR statistic and the residual deviance were chosen, since the LR test is primarily used as a test to detect deviation from the canonical link, while the residual deviance test is used as a goodness of fit test where it is common practice to assume a higher p-value as indication of a good fitting model.

Considering (as it is common practice) a p-value larger than .2 for the LR test as indication of a good fitting model, the test will be unable to detect the large maximal difference of 15% (10%) in probability up to 12.0% (6.4%) in the area of high, up to 28.2% (27.7%) in the area of medium and up to 46.7% (44.8%) in the area of low discrimination power when $n = 20$ ($n = 40$). The percentages are roughly halved when LR test statistics with a p-value greater than .5 are considered.

If one relies only on a residual deviance goodness of fit statistic as measure of goodness assuming a p-value of larger than .5, say, as an indication of a good fitting model, one can see that this test is especially unable to detect link misspecification when $\psi > 1$, i.e. in the area of medium discrimination power. In particular, we observe that up to 8.4% (7.8%) in the area of high, up to 64.9% (44.7%) in the area of medium and 31.4% (36.2%) in the area of low information of the residual deviance test are unable to detect a maximal absolute difference of 15% (10%) in probabilities when $n = 20$ ($n = 40$) assuming a p-value of .5 as indication of a good fitting model. Again, these percentages are roughly halved when a p-value of .75 is assumed as sufficient evidence for the canonical model.

In a second step, we determined the sensitivity of these two tests, i.e we are interested in the number of times the test would reject the canonical model, when in reality there is at most a negligible deviation from the canonical model. For this, we assumed a maximal absolute difference of 5% in probabilities as a negligible deviation from the logistic model. It turns out, that the residual deviance test has less sensitivity against small deviations from the canonical model than the LR test. Both tests, however, are too sensitive in areas of high discrimination power and when the sample size is large ($n = 40$). In particular for the LR test, we observe that for $n = 20$ ($n = 40$) up to 20.8% (34%) in the area of high, up to 15% (23.2%) in the area of medium and up to 9.2% (13%) in the area of low discrimination power to reject the logistic model at $\alpha = .05$ when the true underlying model only deviates by at most 5% in the probabilities from the logistic model. For the residual deviance test, the same percentages are discrimination power when $n = 20$ ($n = 40$).

To summarize, these results clearly demonstrate, that on the one hand there is no guide on how large a p-value has to be, before it gives sufficient indication for a good fitting model. In any case, they have to be much larger than significance levels for rejecting the point null hypothesis. In particular, the residual deviance test turns out to be very poor in detecting a large deviation from the canonical model. In addition, prediction for some covariate values within the range of observed covariate values will be completely unreliable. Therefore, ordinary goodness of fit tests such as the residual deviance test or even the LR test for testing logistic link within the class of generalized logistic links should only be used very carefully to validate the logistic regression model. Large p-values turn out to be misleading. On the other hand, for certain (unknown) parameter configurations which provide high information about the link, both tests are too sensitive to the occurrence of deviations from the logistic model, which are too small to be of importance to the data analyst.

The Pearson χ^2 statistic assuming a logistic model has been also investigated. We obtained similar results for other GLM's which are not displayed by the ease of brevity.

Small Sample Properties of the Validation Test when Verifying the Logistic Model. The same simulation setup has been used to investigate the small sample behavior of the proposed validation test. We consider maximal absolute differences of 15% in probabilities from the logistic model as large deviation, while a maximal absolute difference of 5-10% are tolerable deviations from the link. Therefore, we investigated $H : \Delta_p(\psi) > \Delta_0$ versus $K : \Delta_p(\psi) \leq \Delta_0$ for $\Delta_0 = .1$ and .05. We expect the power of the asymptotic link validation test to be larger for the equally spaced covariate case compared to the unequally spaced case and when the true link has a heavier right tail ($\psi < 1$) compared to a lighter right tail ($\psi > 1$). Again sample sizes of n=20 (solid line) and n=40 (dotted line) were studied. Cases where the maximization routine failed to converge were deleted. The observed power based on 500 binomial data sets for $\Delta_0 = .1$ is presented in Figure A.3.

In all cases considered, the validation test allowing for 10% maximum absolute difference in probabilities maintains its significance level α of .1, with being more conservative on the left hand side of the alternative K ($\psi = .35$) and more liberal on the right hand side ($\psi = 1.84$). A possible explanation for this is the smaller area of large difference between the link parameters for $\psi > 1$ compared to $\psi < 1$. For the same reason, the power of the test is higher by about 50% for the equally spaced covariate case. The power of the test increases by about 50% as sample size changes from $n = 20$ to $n = 40$.

The reduction in power is large especially for $n = 20$ and unequally spaced covariates. Even for $n = 40$ the maximal power is .3, indicating that larger sample sizes than 40 are required. However, the test maintains equally well its significance level of $\alpha = .1$ at both end points of the alternative ($\psi = .65$ and $\psi = 1.39$).

In summary, the validation test maintains its significance level. The power of the test depends on whether data is collected in areas of large difference between the logistic link and the generalized logistic link. For a validation neighborhood of 10% in probabilities a sample size of $n = 20$ is sufficient for a maximal power of .5 in the case when the data can determine the areas of the large difference while a sample size of $n = 40$ is needed for data which is sparse in areas of large difference. For the smaller validation neighborhood of 5% in probabilities, sample sizes larger than $n = 40$ are required. Hence, the proposed classification of regions of the parameter space into different zones of discrimination power is an excellent indicator for the actual power of the validation and discrimination test.

For Poisson responses we use the following link family (Czado 1992)

$$F(\eta, \psi) = \exp(h(\eta, \psi)), \tag{17}$$

where $h(\eta, \psi)$ was previously defined by (2). $\psi = 1$ corresponds to the canonical link. We will denote these links as generalized Poisson links. We investigate now in detail the behavior of the link validation test based on the maximal change in the no event probabilities $\Delta_{np}(\psi)$ defined in (15). A surface plot (not shown) of the absolute difference in the no event probabilities for the generalized Poisson link model and the ordinary Poisson regression model reveals the same features as shown in Figure 3.1 for the binomial responses. Therefore, we again can classify four areas of varying degree of information about the link parameter ψ, namely area of high discrimination power when $\psi < 1$ and a large range of linear predictors η_i, area of low discrimination power when $\psi, 1$ and a small range of η_i's and medium discrimination power when $\psi > 1$. Table A.1 gives for some special Δ_0 values the corresponding interval $[\psi_* - \psi_l, \psi_* + \psi_u]$ to insure $\Delta_{np}(\psi) < \Delta_0$ for all $\psi \in [\psi_* - \psi_l, \psi_* + \psi_u]$:

	$\Delta_{np}(\psi) = \Delta_0$						
	.01	.025	.05	.075	.1	.15	.20
$\psi_* - \psi_l$		.65	.32	-.001	-.32	-1.00	-1.80
$\psi_* + \psi_u$		1.37	1.79	2.26	2.80	4.22	6.49

Table A.1: Choice of the link parameter ψ to achieve a maximal absolute difference $\Delta_{np}(\psi) = \Delta_0$ between the generalized Poisson model and canonical Poisson model

Data from the following generalized Poisson model with a single covariate was generated

$$Y_i \sim \text{ Poisson}(\mu_i) i = 1, \cdots, k \text{ with } \mu_i = F(\beta_0 + \beta_1 x_i, \psi)$$

where $F(\eta, \psi)$ is given by (17) and $\beta_0 = -1$. We present only an equally spaced covariate case, where $\mathbf{x} = (\mathbf{x_1}, \cdots, \mathbf{x_n})$ is equally spaced between .5 and 2.5. Three values for the regression slope β_1 were chosen: $\beta_1 = 1.5, 2, 2.5$. This allows for a maximal value of η_i of 2.75,4,5.25, respectively, i.e. for $\psi < 1$ we expect lowest (highest) power of the link validation test for $\beta_1 = 1.5 (\beta_1 = 2.5)$. Finally, we investigated $n = 25$ and $n = 50$. Asymptotic validation tests of $H : \Delta_{np}(\psi) > .05$ versus $K : \Delta_{np}(\psi) \leq .05$ were performed at level $\alpha = .1$ using 500 Poisson regression data sets and the observed power is given in Figure A.4. Cases where the maximization routine failed to converge were deleted.

As for binomial responses, the validation test allowing for 5% maximum absolute difference in no event probabilities maintains its significance level α of .1 with being more conservative on the left hand side of the alternative $K(\psi = .32)$ and more liberal on the right hand side ($\psi = 1.79$). A similar explanation for this as in the binomial case can be given. As expected the power increases as the range of η_i increases. Further, there is no substantial increase in power when we increase n from 25 to 50.

In summary, the asymptotic link validation test performs very well for Poisson responses; it maintains its significance level and the power depends on whether data is collected in areas of large difference between the canonical link and the generalized Poisson link. A sample size of $n = 25$ is sufficient.

Appendix B (Proofs)

Fahrmeir and Kaufmann (1985) proved the asymptotic results for ordinary GLM's, i.e for GLM's with fixed link. Their results are now extended to the case of an estimated link parameter ψ.

(1985) for noncanonical links (see Section 4.1) can be followed using a Taylor expansion of the log likelihood $l(\delta)$ around $\boldsymbol{\delta}_0$.

Sketch of proof for Theorem 2.2. First, an analogue of Lemma 2 (Fahrmeir and Kaufmann (1985), p. 361) will be derived:

Lemma B1 *Under (R1) and (R3),* $I_n s_n \xrightarrow{\mathcal{D}} N_{p+1}(0, I)$ *as* $n \to \infty$.

Proof of Lemma B1. As in Fahrmeir and Kaufmann (1985), the proof uses the central limit theorem for triangular arrays and establishes the validity of the Lindeberg condition. For this, define the triangular array

$$Z_{ni} = \lambda^t I_n^{-\frac{1}{2}} \boldsymbol{s}(y_i, x_i, \boldsymbol{\delta}_0)$$

where $\boldsymbol{s}(y_i, x_i, \boldsymbol{\delta}_0)$ is the vector of individual score contributions, i.e. given by:

$$s(y_i, x_i, \boldsymbol{\delta}_0) = a(\phi)^{-1}(d_i x_{i1} F_{i1}(y_i - \mu_i), \cdots, d_i x_{ip} F_{i1}(y_i - \mu_i), d_i F_{i2}(y_i - \mu_i))$$

Define $\alpha_{ni} = \lambda^t I_n^{-\frac{1}{2}} \mathbf{L}(x_i^t \beta_0, \psi_0)$, where

$$\mathbf{L}(x_i^t \beta_0, \psi_0) = a(\phi)^{-1}(d_i x_{i1} F_{i1}, \cdots, d_i x_{ip} F_{i1}, d_i F_{i2}).$$

Note that this vector above is bounded when $\{x_n, n \geq 1\}$ by condition (R3), since d_i is a continuous function of $F(x_i^t \beta_0, \psi_0)$. We can now express Z_{ni} as

$$Z_{ni} = \alpha_{ni}(Y_i - F(x_i^t \beta_0, \psi_0)).$$

Under (R1) and (R3), we have with the Cauchy-Schwarz inequality

$$max_{i \leq n} \alpha_{ni}^2 \leq ||\mathbf{L}(x_i^t \beta_0, \psi_0)||^2 \lambda_{min} I_n^{-1} \leq k \lambda_{min} I_n^{-1} \to 0 \text{ as } n \to \infty.$$

Then we argue as in Fahrmeir and Kaufmann (1985,p363) for compact regressors that the Lindeberg condition is satisfied.

Sketch of proof for Theorem 2.1. Using Lemma 3.1 and the following condition

R5 For every $\varepsilon > 0$

$$max_{\delta \in N_n(\varepsilon)} ||V_n(\boldsymbol{\delta}) - I|| \to 0 \text{ in probability}$$

where $V_n(\delta) = I_n^{-\frac{1}{2}} H_n(\delta) I_n^{-\frac{t}{2}}$ and $N_n(\varepsilon)$ is defined as $N_n(\varepsilon) = \{\delta : ||I_n^{\frac{t}{2}}(\boldsymbol{\delta} - \boldsymbol{\delta}_0)|| \leq \varepsilon\}$,

we proceed as for the proof of Theorem 3 in Fahrmeir and Kaufmann (1985). Finally, it remains to show that (R1),(R3) and (R4) are sufficient for (R5). For this, we argue as in the proof of Theorem 4 (Fahrmeir and Kaufmann, p.364) considering the same partition of matrices as for $I_n(\delta)$ to adjust for the additional estimation of the link parameter ψ.

Proof of Lemma 5.1. It is sufficient to show that $||X^t D(\mu, a) X||$ is strictly decreasing. To this end let $0 \leq a_1 < a_2$ and rewrite $D(\mu, a_1) = A(\mu) + D(\mu, a_2)$ where

$$A(\mu) = diag\left\{\frac{(a_1 - a_2)\,\mu_i^2}{(1 + a_1\mu_i)(1 + a_2\mu_i)}\right\}$$

Now apply the spectral decomposition theorem and the fact that the set of positive definite matrices is closed under multiplication and forming inverses. This shows that $||A(\mu)D^{-1}(\mu, a_2) + I|| < 1$, where I denotes the unity matrix.

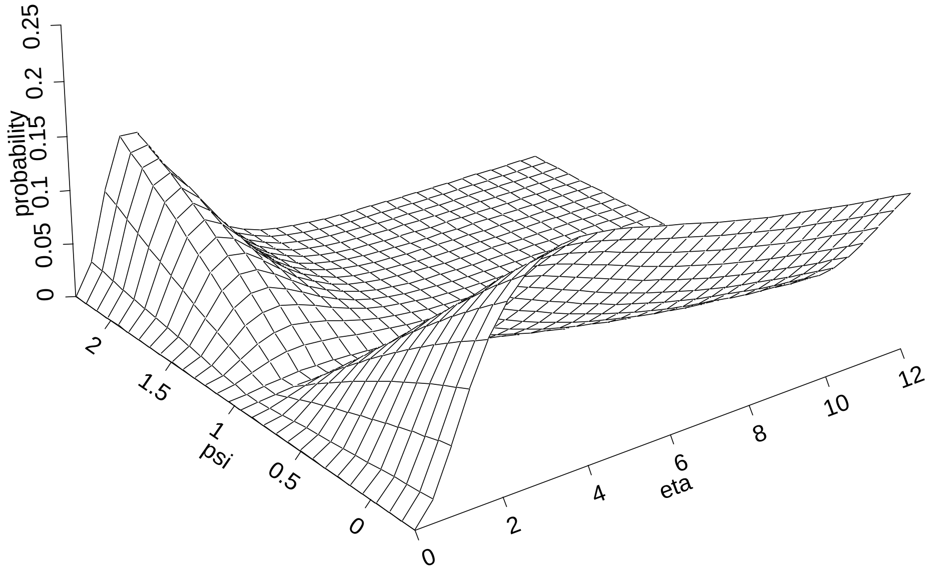

Figure 3.1: Absolute difference in probabilities between the logistic and generalized logistic distribution

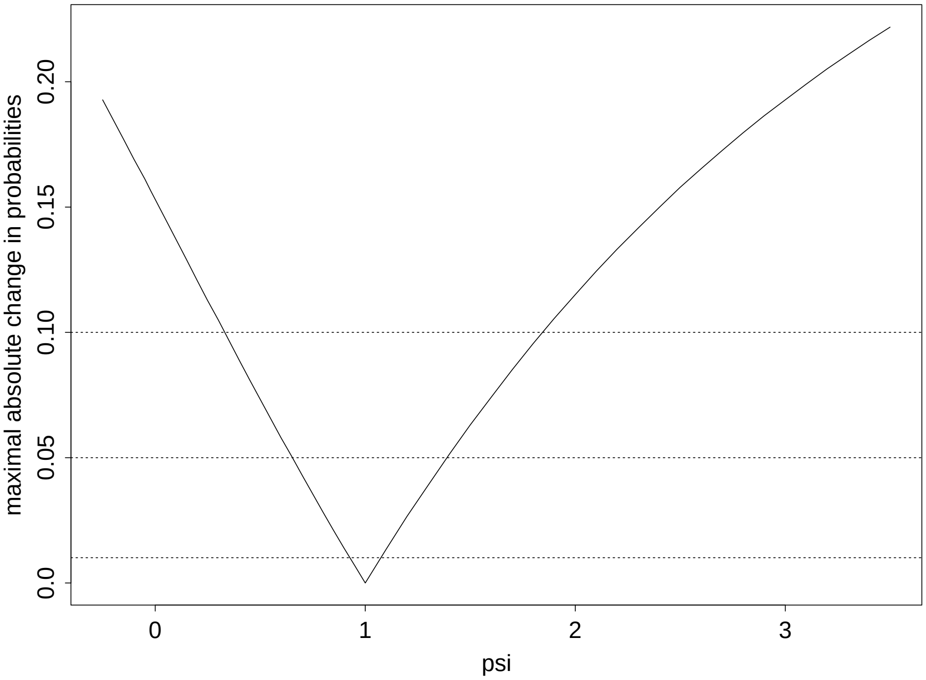

Figure 3.2: Maximal absolute difference $(\Delta_p(\psi))$ between the logistic and the generalized logistic distribution

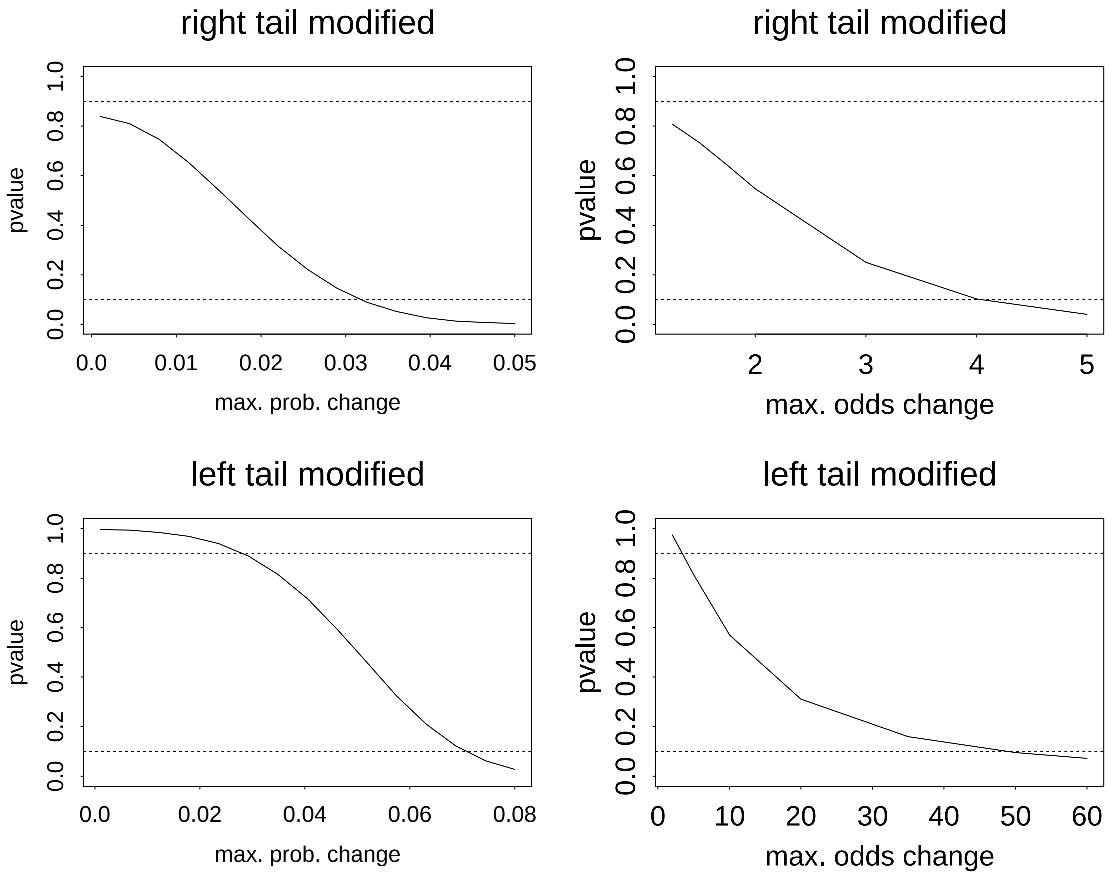

Figure 4.1: P-value curves for assessing the maximal change to estimated success probabilities and odds for the Age to Menarche Data

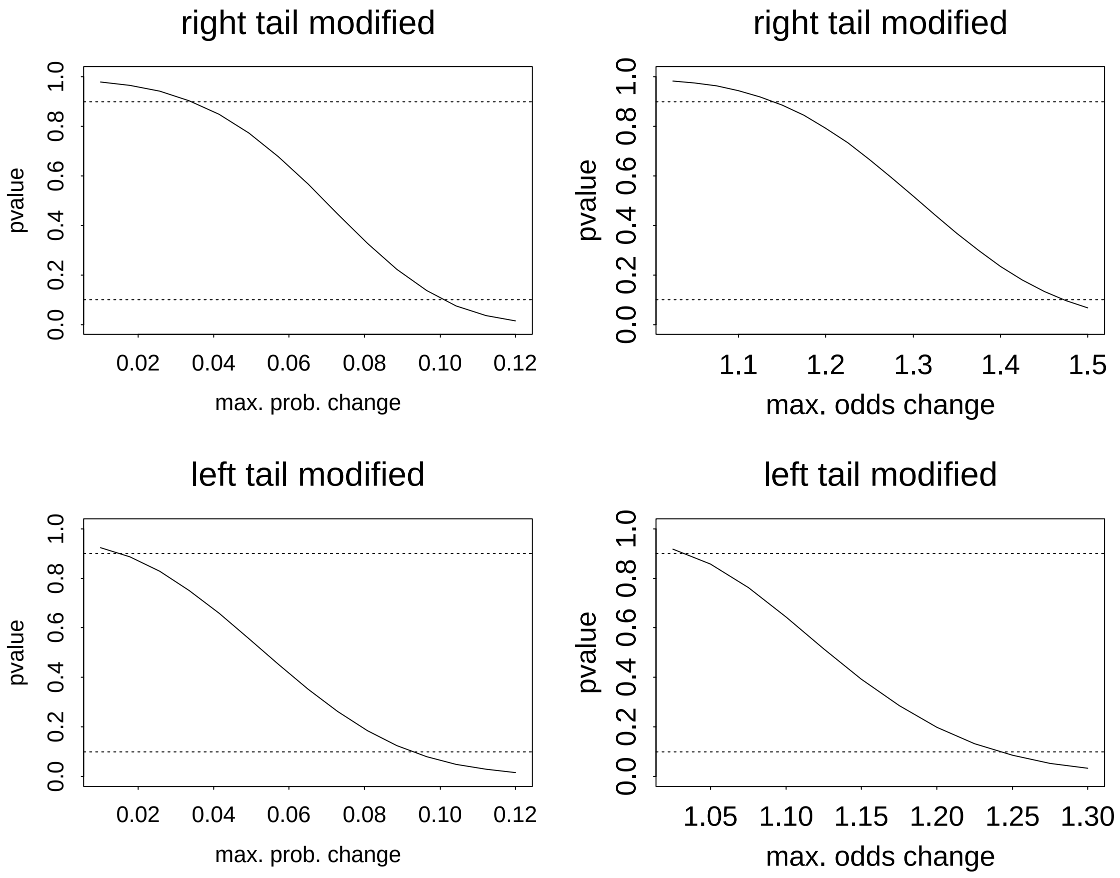

Figure 4.2: P-value curves for assessing the maximal change to estimated success probabilities and odds for the Bottle Deposit Data

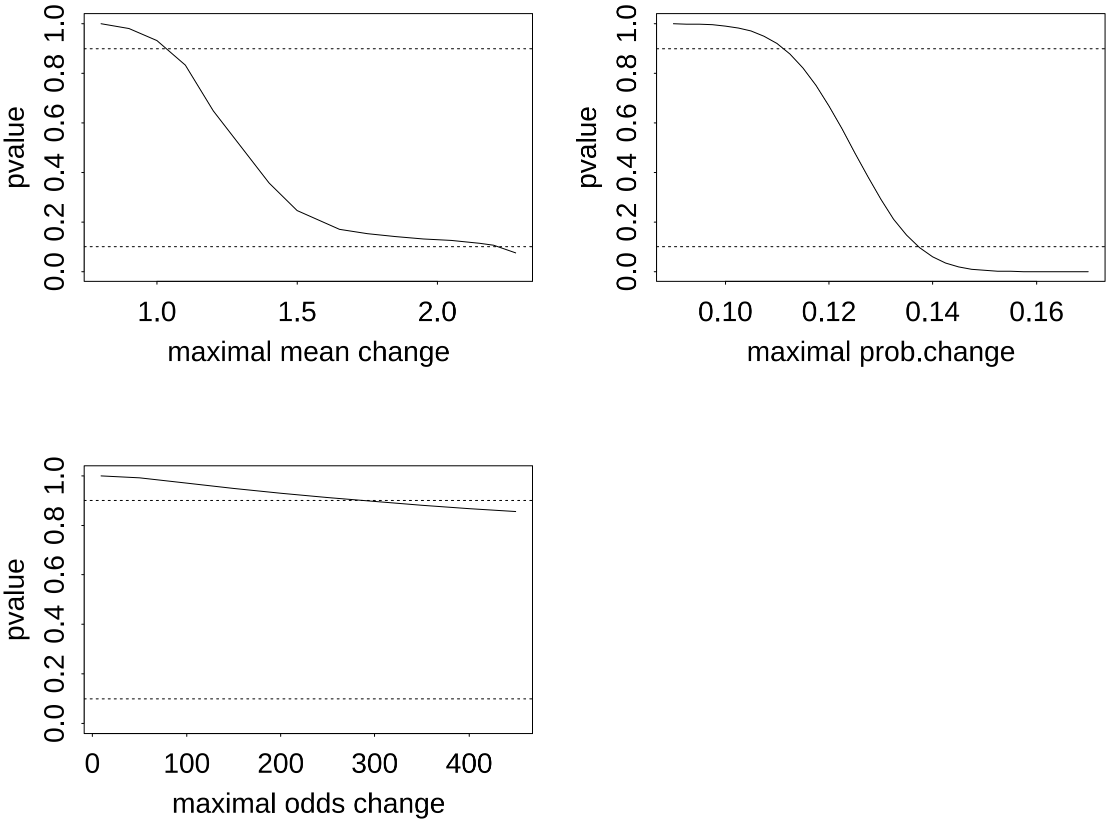

Figure 4.3: P-value curves for assessing the maximal change to estimated mean responses, no event probabilities and odds for the no event probabilities for the Mining Fracture Data

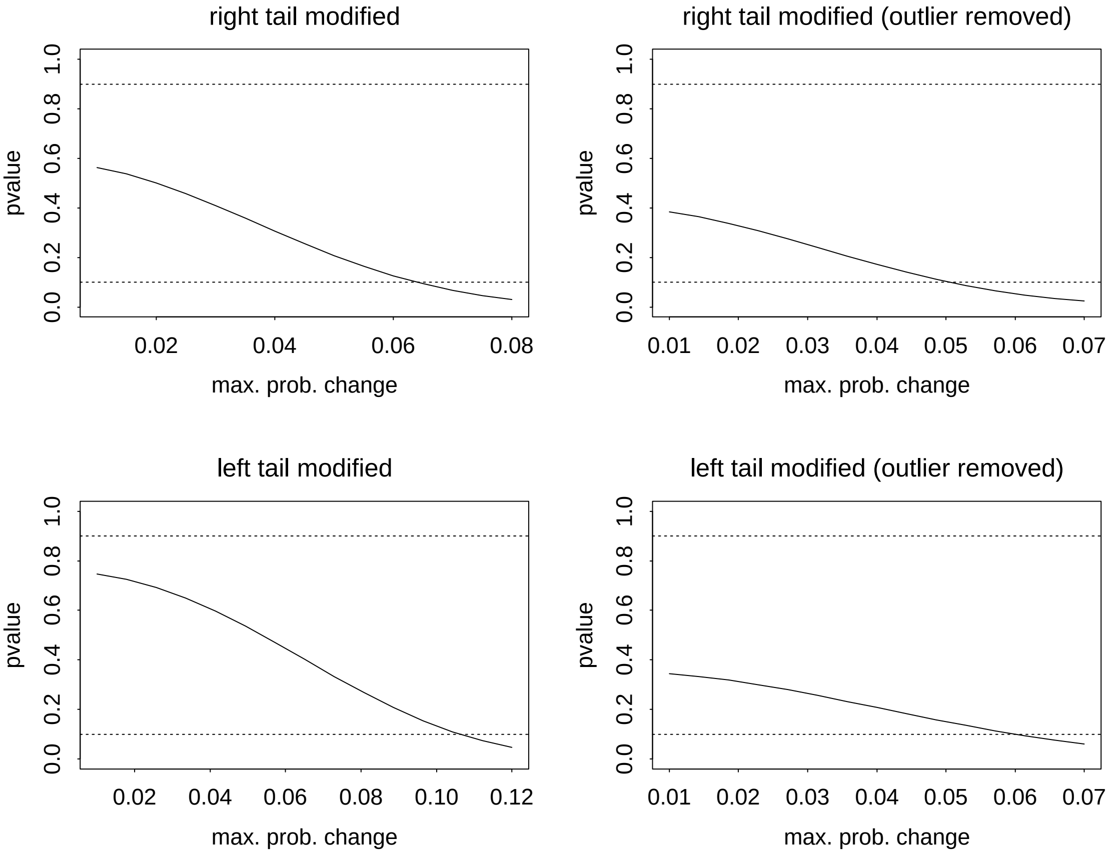

Figure 5.1: P-value curves for assessing the maximal change to estimat ed success probabilities for the complete and outlier removed Flour Beetle Data

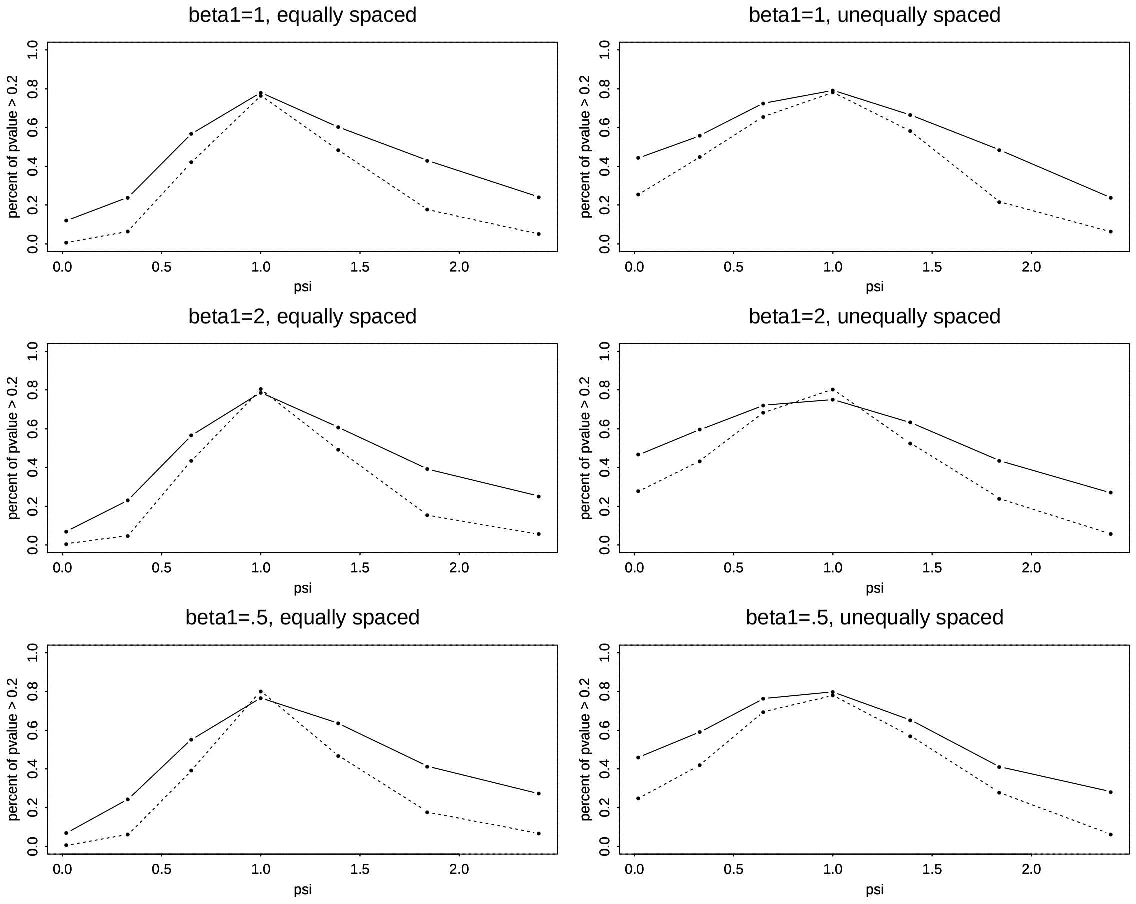

Figure A.1: Percentage of likelihood ratio tests of $H : \psi = 1$ versus $K : \psi \neq 1$ which result in a p-value $> .2$

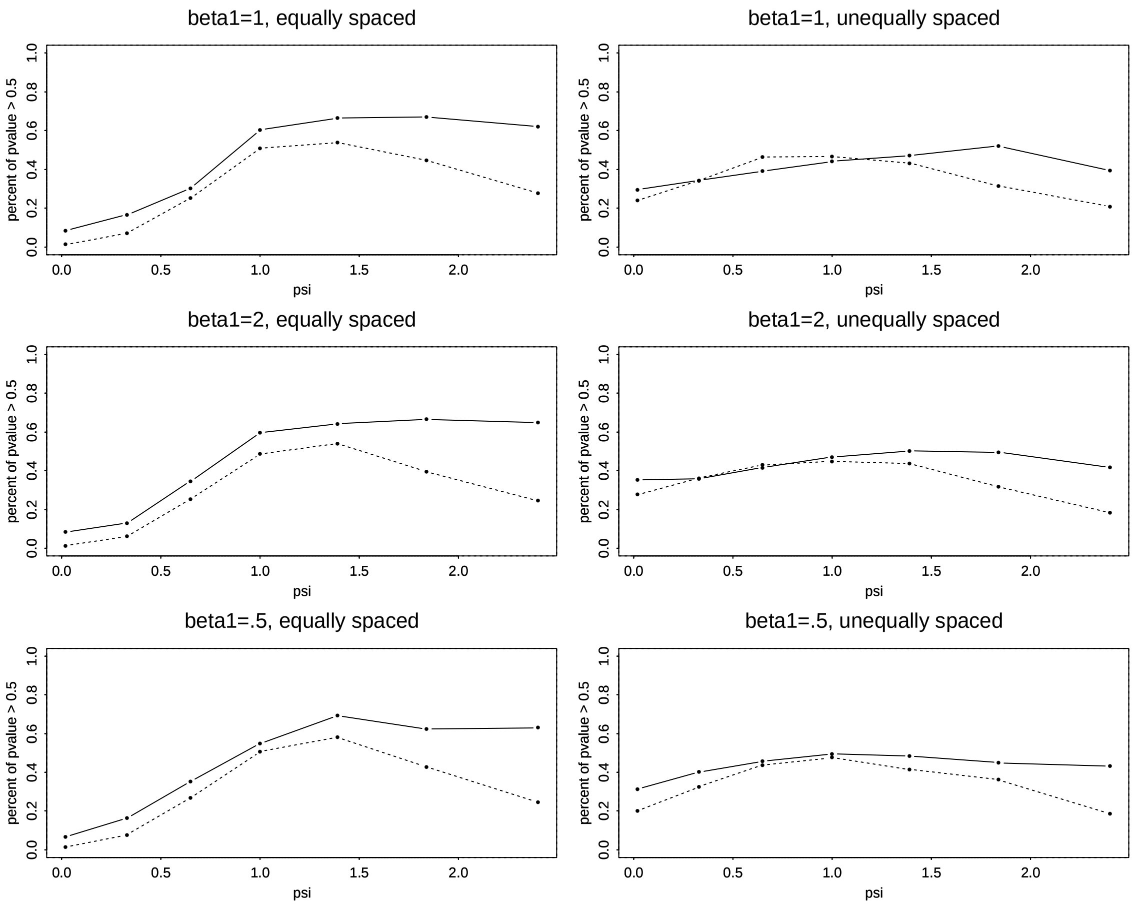

Figure A.2: Percentage of residual deviance tests assuming a logistic model which result in a p-value $> .5$

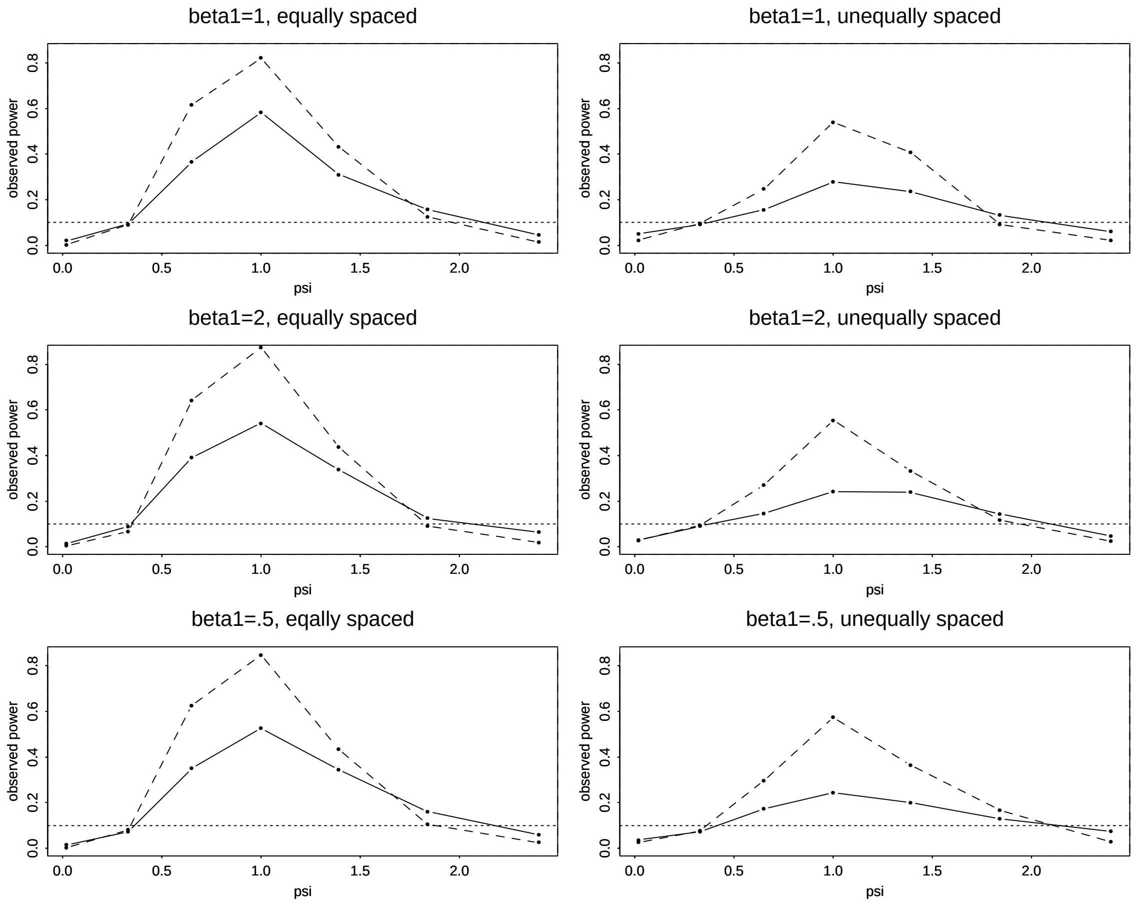

Figure A.3: Observed power of the link validation test for binomial responses of $H : \Delta_p(\psi) > .1$ versus $K : \Delta_p(\psi) \leq .1$ at significance level $\alpha = .1$

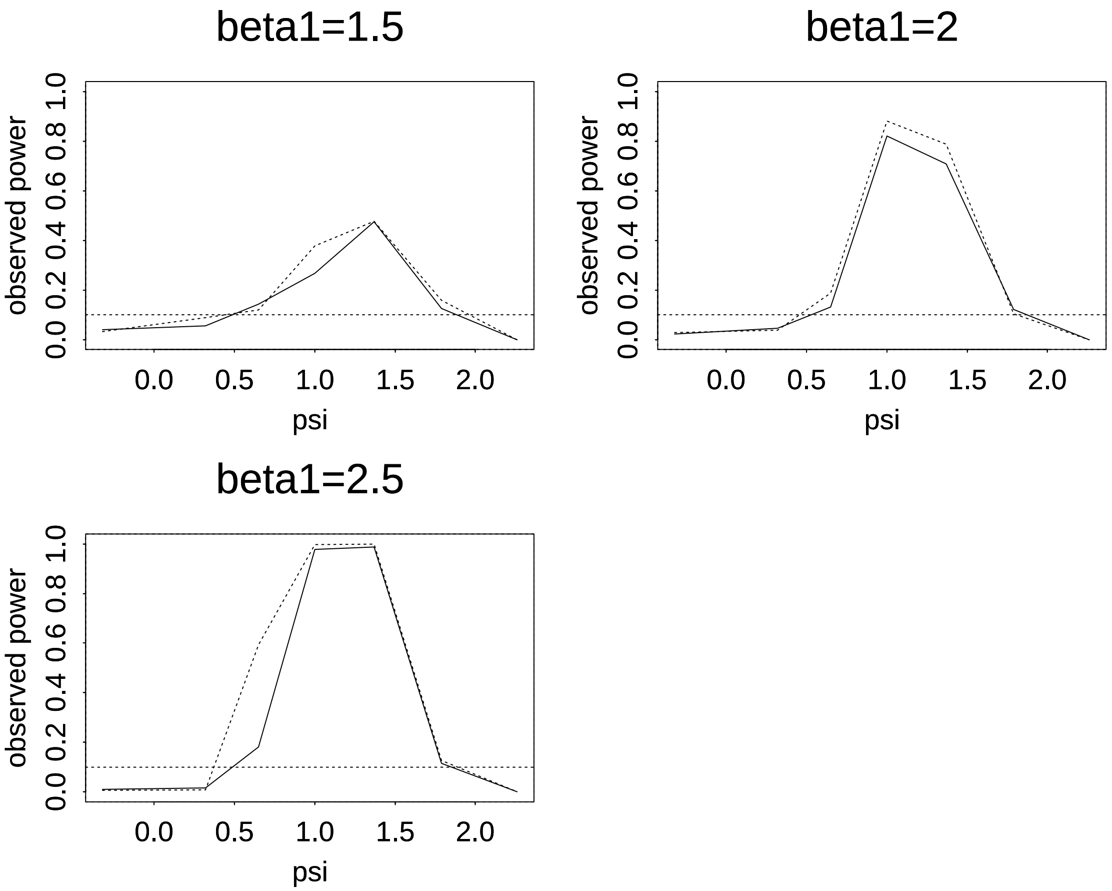

Figure A.4: Observed power of the link validation test for Poisson responses of $H : \Delta_{np}(\psi) > .05$ versus $K : \Delta_{np}(\psi) \leq .1$ at significance level $\alpha = .05$ (n=25(50) solid (dotted) line)