Czado, Sikora:

Quantifying overdispersion effects in count regression data

Sonderforschungsbereich 386, Paper 289 (2002)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Quantifying overdispersion effects in count regression data

Claudia Czado and Inka Sikora
Zentrum Mathematik
Technische Universität München
Boltzmannstr. 3
D-85747 Garching bei München, Germany
cczado@ma.tum.de, inka@ma.tum.de

Abstract

The Poisson regression model is often used as a first model for count data with covariates. Since this model is a GLM with canonical link, regression parameters can be easily fitted using standard software. However the model requires equidispersion, which might not be valid for the data set under consideration. There have been many models proposed in the literature to allow for overdispersion. One such model is the negative binomial regression model. In addition, score tests have been commonly used to detect overdispersion in the data. However these tests do not allow to quantify the effects of overdispersion. In this paper we propose easily interpretable discrepancy measures which allow to quantify the overdispersion effects when comparing a negative binomial regression to Poisson regression. We propose asymptotic α-level tests for testing the size of overdispersion effects in terms of the developed discrepancy measures. A graphical display of p values curves can then be used to allow for an exact quantification of the overdispersion effects. This can lead to a validation of the Poisson regression or a discrimination of the Poisson regression with respect to the negative binomial regression. The proposed asymptotic tests are investigated in small samples using simulation and applied to two examples.

Key words: Poisson regression, negative binomial regression, quantification of overdispersion.

1 Introduction

In many areas of application, like automobile insurance, life insurance, biostatistics or physical applications, Poisson regression is used to model the dependency of count data on covariates. Since the Poisson regression model is a generalized link model (GLM) with canonical link, the regression parameter can be easily estimated. One restriction of the Poisson distribution is however that it allows only a single parameter to estimate the mean and variance. In particular equidispersion, i.e. equality of mean and variance, holds. If the variance exceeds the mean we speak of overdispersion.

There is considerable literature on modeling and detecting overdispersion in count regression data. Roughly, two general approaches are followed. One is based on mixture models, where random effects are included in the Poisson regression to account for overdispersion, while the other approach is to develop probability models for count data which have more than one parameter for the modeling of mean and variance.

Mixture models are mostly used for the detection of overdispersion. In particular score tests are derived, which have the advantage that only the model without the random effects has to be fitted to the data. Dean (1992) provides a unifying theory for score tests for extra Poisson variation developed by Fisher (1950), Collings and Margolin (1985), Cameron and Trivedi (1986) and Dean and Lawless (1989). These tests assume more general exponential family models with random effects and are based on expansions developed by Cox (1983) and Chesher (1984). General exponential mixing models are introduced and discussed by Lindsay (1986). Score tests for extra Poisson variation in the positive or truncated-at-zero Poisson regression model against truncated-at-zero negative binomial family alternatives are derived in Gurmu (1991) and score tests of extra Poisson variation in left or right truncated Poisson regression models are given in Gurmu and Trivedi (1992). Graphical methods for the detection of overdispersion are developed in Lambert and Reoder (1995). The special case of finite mixed Poisson regression models with covariates in both Poisson rates and mixing probabilities is discussed in Wang, Cockburn, and Puterman (1998).

Breslow and Clayton (1993) use penalized quasi-likelihood and McCulloch (1997) proposes a Monte Carlo EM algorithm for estimation in generalized linear mixed models (GLMM). These include Poisson regression with random effects. Markov Chain Monte Carlo (MCMC) methods are utilized in Bayesian analyses for GLMM's (see for example Zeger and Karim 1991, Besag, York, and Mollie 1991 or Clayton 1996).

The second approach to modeling count regression data with overdispersion uses probability models for count data with more than one parameter. One of the most known such model is the negative binomial regression model (see for example Lawless 1987). Note that the negative binomial regression model can also be derived as Poisson-Gamma mixture model (see for example Cameron and Trivedi 1998, p. 101). Other count probability models are the double Poisson family, which is a special case of the double exponential models (see Efron 1986) and the generalized Poisson distribution (Consul 1989). Inference in the double exponential regression models is only approximate. Ganio and Schafer (1992) develop likelihood ratio and score tests for testing for overdispersion in a double exponential family, while Fitzmaurice (1997) considers the problem of model selection for overdispersed data within the class of double exponential family models. In contrast to the double exponential families, Consul and Famoye (1992) give maximum likelihood estimates in generalized Poisson regression models and an application using the generalized Poisson regression model is given in Singh and Famoye (1993).

In contrast to methods for detecting overdispersion in a data set, the goal of this paper is to develop quantitative measures for overdispersion, which can be easily interpreted and assessed by the experimenter. These measures will be used to quantify the effects of overdispersion in a data set. In particular we seek to answer the following question:

> *Is overdispersion in the data low enough to ignore it by using the Poisson regression, or is the effort justified, to switch from the Poisson model to the negative binomial regression model because of a high degree of overdispersion in the data?*

Score tests for overdispersion are typically based on hypotheses of the following form

$$H : a = a_0 \text{ versus } K : a > a_0, \tag{1.1}$$

where a is the overdispersion parameter and $a = a_0$ corresponds to no overdispersion. In many situations $a_0 = 0$. We like to point out that the above question cannot be answered by a test for (1.1), since the rejection of H does not allow for a quantification of overdispersion. A more appropriate way to answer the above question is to consider the following test problem

$$H : a > a_0 \text{ versus } K : a \leq a_0, \tag{1.2}$$

where the bound a_0 is chosen in advance. If the null hypothesis of test problem (1.2) is rejected we have significant evidence that the overdispersion parameter a is bounded by

a_0. Thus a quantification of the overdispersion in terms of the size of the overdispersion parameter a is achieved. Similar approaches are common in bioequivalence testing (see for example Chow and Liu (1992) for normal populations and Munk and Czado (1998) for a nonparametric approach). This approach has also been successfully used in the quantification of link misspecifications in GLM's (see Czado and Munk (2000)).

We will first develop an asymptotic α level test for testing problem (1.2), however since the size of the dispersion parameter a in the negative binomial model is difficult to interpret, we develop more interpretable measures depending on a. These measures are also suitable for detecting the effects of outliers on the degree of overdispersion. The statistical test for (1.2) will then be used to construct a corresponding test for the more interpretable measures.

The small sample performance of the constructed asymptotic α level test will be investigated in a comprehensive simulation study (see Appendix), which leads to a modification of the original rejection area. Finally, we demonstrate the usefulness of these modified test procedures in two data sets. The first data set involves claims from an automobile insurance. Here we can show that while a negative binomial regression model fit gives a very small value for the overdispersion parameter, strong overdispersion effects are present when outliers are not accounted for. In a second example involving patent data strong overdispersion effects are visible even after accounting for possible outliers.

The paper is organized as follows: Regression models for count regression data including GLM's and negative binomial regression are introduced and discussed in Section 2. This includes also the derivation of an asymptotic α level test for (1.2) for the overdispersion parameter in a negative binomial regression setup. Section 3 develops more interpretable discrepancy measures and uses p-value curves to allow for an exact quantification of overdispersion. Examples are given in Section 4 and Section 5 summarizes and discusses the results. Finally in the appendix the results of a simulation study investigating the finite sample properties of the asymptotic test procedure developed in Section 2 are given.

2 Regression models for count regression data

In the last few years usage of GLM's has become standard in many areas of application. The standard reference on GLM's is McCullagh and Nelder (1989) and further details, especially for count data, can be found in Cameron and Trivedi (1998). We now sketch the required set up and provide some details especially for the modeling of count regression data. In general

there are 3 components needed to define a GLM:

1. Random component:
   Given the covariates $\boldsymbol{x_i}$, the responses $y_i,\ i = 1, .., n$ are iid with a density of the form
   $$f(y; \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{2.1}$$
   which is a density from the exponential family.

2. Systematic component:
   The given p vectors of covariates $\boldsymbol{x_j}$, $j = 1, .., p$ define linear predictors
   $\eta_i = \boldsymbol{x_i^t \beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + .. + \beta_p x_{ip}$ for $i = 1, .., n$, where $\boldsymbol{\beta}$ is the vector of the unknown regression parameters.

3. Parametric link component:
   The link function $g(\mu_i) = \eta_i = \boldsymbol{x_i^t \beta}$ combines the linear predictor with the mean μ_i of y_i. Here, the canonical link function is used, so that $\boldsymbol{\theta} = \boldsymbol{\eta}$ holds.

As a first model for count regression data we consider Poisson regression.

2.1 Poisson Regression

The density of the Poisson distribution $f_{poi}(y) = \exp\{-\mu + y \ln \mu - \ln y!\}$ can be rewritten in the form of the exponential family with $a(\phi) = 1,\ \theta = \ln \mu,\ b(\theta) = \exp(\theta) = \exp(\ln \mu) = \mu$, and $c(y, \phi) = -\ln y!$. The canonical link function is the log-link function $\eta = \ln(\mu)$. Therefore, the conditional mean of the observations y_i given the covariates $\boldsymbol{x_i}$ is given by

$$E(Y_i \mid \boldsymbol{x_i}) = \exp(\eta_i) = \exp(\boldsymbol{x_i^t \beta}),\ i = 1, .., n.$$

One important property of the Poisson distribution is equidispersion, i.e. $\mu_i = E(Y_i) = Var(Y_i)$. If in a data set the variance exceeds the mean, we say, that the data shows overdispersion. We speak of underdispersion, if the variance is lower than the mean.
Usually, we assume, that the observation periods for count data Y_i are identical for all $i = 1, .., n$. But observation periods can be different in practice. To standardize the observation time for all observations an additional variable t_i must be introduced, which is related to the observation period of the i-th observation. It is now assumed, that the response $Y_i \mid \boldsymbol{x_i}$ is Poisson distributed with parameter $t_i \mu_i^*$, where $\mu_i^* = \exp(\boldsymbol{x_i^t \beta})$. This implies that

$$E(Y_i \mid \boldsymbol{x_i}) = t_i \mu_i^* = \exp(\ln(t_i) + \boldsymbol{x_i^t \beta}),$$

and the linear predictors are given by

$$\eta_i = \boldsymbol{x_i^t}\boldsymbol{\beta} = \ln(\mu_i) - \ln(t_i) = \theta_i - \ln(t_i),$$

where $\ln(t_i)$ is called offset and is assumed to be known.

As already discussed in the introduction there are different possibilities to construct overdispersion models. We concentrate in this paper on the negative binomial regression model, which we discuss now.

2.2 Negative binomial regression

One of the earliest references on the negative binomial distribution is Greenwood and Yule (1920). Properties of the negative binomial distribution are discussed in Armitage and Colton (1998, p. 2962-2967) and in Johnson, Kotz, and Kemp (1993, p.99). The density of the negative binomial distribution is defined as follows:

$$f(y) = \frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!}\left(\frac{\mu}{a^{-1}+\mu}\right)^y\left(\frac{a^{-1}}{a^{-1}+\mu}\right)^{a^{-1}} \tag{2.2}$$

with dispersion parameter $a \geq 0$. Clark and Perry (1989) use quasi maximum likelihood for parameter estimation and Aragón, Eberly, and Eberly (1992) discuss the existence and uniqueness of the maximum likelihood parameter estimates in the negative binomial distribution.

In contrast to the Poisson distribution the negative binomial distribution allows overdispersion, since the mean, denoted by $\mu_i := E(Y_i)$, is always lower than the variance, given by $Var(Y_i) = \mu_i + a\mu_i^2$, if $a > 0$. For example, Cameron and Trivedi (1998, p.75) show, that the Poisson distribution is a special case of the negative binomial distribution, when the dispersion parameter a is equal to zero. For known a it is straight forward to see that (2.2) is a member of the exponential family given in (2.1). For an unknown parameter a Cameron and Trivedi (1998, p.33 and p.75), show that the negative binomial distribution is a member of the exponential family with nuisance parameter ϕ, which is similar to the exponential family defined in (2.1). The Fisher information matrix of the negative binomial model is given in Lawless (1987). To get the entries of the Fisher information matrix, one has first to take the second partial derivatives of the log-likelihood function, which is given by

$$\mathcal{L}(\beta, a) = \ln L(\boldsymbol{\beta}, a) = \sum_{i=1}^{n}\left[\sum_{j=0}^{y_i^*}\ln(aj+1) - \ln(y!) + y_i \ln\mu_i - (y_i + a^{-1})\ln(1+a\mu_i)\right],$$

where $y_i^* = y_i - 1$, and then take the expectations of the negative second derivatives. It follows, that the expectation of the second mixed derivatives is zero, which implies, that the parameters a and $\boldsymbol{\beta}$ are asymptotically uncorrelated and the Fisher information matrix has the following form of a diagonal block matrix:

$$FI(\boldsymbol{\beta}, a) = \begin{bmatrix} FI_{r,s}(\boldsymbol{\beta}, a) & \mathbf{0} \\ \mathbf{0}^t & FI_{p+1,p+1}(\boldsymbol{\beta}, a) \end{bmatrix} \qquad \text{for} \quad r, s = 1, ..., p.$$

Here $FI_{r,s}(\boldsymbol{\beta}, a)$ is a $p \times p$ matrix, while $FI_{p+1,p+1}(\boldsymbol{\beta}, a)$ is a scalar. It is not trivial to obtain the joint MLE $(\widehat{\boldsymbol{\beta}}, \widehat{a})$ of $(\boldsymbol{\beta}, a)$. The simplest method to get the estimator $(\widehat{\boldsymbol{\beta}}, \widehat{a})$ is to maximize the log-likelihood function $\mathcal{L}(\boldsymbol{\beta}, a)$ for fixed values a, for example with the Fisher-scoring or Newton-Raphson algorithm. From the resulting estimates $\widehat{\boldsymbol{\beta}}(a)$ and the profile likelihood $\mathcal{L}(\widehat{\boldsymbol{\beta}}(a), a)$, which still depends on a, one can get the estimator $\widehat{a}$ with one-dimensional maximization with respect to a.

2.3 Asymptotic $\alpha-$ level test for the testing problem (1.2) in negative binomial regression

Lawless (1987) shows, that under regularity conditions and mild conditions on the covariate values $\boldsymbol{x_i}$'s to ensure that $n^{-1}F\ I(\boldsymbol{\beta}, a)$ approaches a positive definite limit as $n \to \infty$, the MLE $(\widehat{\boldsymbol{\beta}}, \widehat{a})$ is asymptotically normally distributed, i.e.

$$FI(\widehat{\boldsymbol{\beta}}, \widehat{a})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \widehat{a} - a) \xrightarrow{\mathcal{D}} N_{p+1}(\mathbf{0}, I_{p+1}). \tag{2.3}$$

Here $N_p(\mu, \Sigma)$ denotes a p-dimensional normal distribution with mean vector μ and covariance matrix Σ, and I_{p+1} denotes the $(p+1)$-dimensional unity matrix. We will now use the asymptotic normality result (2.3) to construct an asymptotic validation test for overdispersion.

Theorem 2.1 *Under the assumption, that (2.3) holds, an asymptotic uniformly most powerful (UMP) α-level test for the hypothesis $H : a > a_0$ versus the alternative $K : a \leq a_0$ is given by the rejection region*

$$\mathcal{C} = \left\{ \widehat{a} : \Phi\left(\frac{\widehat{a} - a_0}{\widehat{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{a})} \right) \leq \alpha \right\}, \tag{2.4}$$

where $\widehat{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{a})$ denotes the $(p+1, p+1)$-entry in the inverse of the Fisher information matrix and $\Phi(.)$ the standard normal cdf.

Proof: *To use the Theorem 3.3.2 (p.78) in Lehmann (1986), we show, that the assumptions,*

needed for this theorem, are satisfied asymptotically. At first one has to prove, that there is asymptotically a monotone increasing likelihood ratio in $\widehat{a}$. Let $p_a(\widehat{a})$ denote the density of $\widehat{a}$ when the true dispersion parameter is a. For this, we choose $a_1 > a_2$. Using (2.3) it follows that

$$\begin{aligned}\frac{p_{a_1}(\widehat{a})}{p_{a_2}(\widehat{a})} &\approx \frac{\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\exp\{-\frac{1}{2\widehat{\sigma}^2}(\widehat{a}-a_1)^2\}}{\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\exp\{-\frac{1}{2\widehat{\sigma}^2}(\widehat{a}-a_2)^2\}} \\ &= \exp\left\{\frac{1}{2\widehat{\sigma}^2}\left(2\widehat{a}(a_1-a_2)+a_2^2-a_1^2\right)\right\},\end{aligned}$$

where $\widehat{\sigma} = \widehat{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{a})$. Since $a_1 - a_2$ is positive, the resulting likelihood ratio is a monotone increasing function in $\widehat{a}$. Now, Theorem 3.3.2, p.78 in Lehmann (1986) can be used asymptotically. Since Lehmann gives a UMP-test for $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$, we have to consider the dual test problem. The first part of the theorem can be transformed to
(i^*): For the test $H : \theta > \theta_0$ versus $K : \theta \leq \theta_0$ there exists a UMP-test, which is given by

$$\Psi(x) = \begin{cases} 1 & \text{if } T(x) < C, \\ \gamma & \text{if } T(x) = C, \\ 0 & \text{if } T(x) > C, \end{cases}$$

where C and γ are determined by $E_{\theta_0}\Phi(x) = \alpha$.
With these results the rejection region for the asymptotic validation test $H : a > a_0$ versus $K : a \leq a_0$ is given by

$$\mathcal{C} = \{\widehat{a} : \ \widehat{a} < C\},$$

where C is uniquely determined by

$$\widehat{P}(C, a_0, \widehat{\sigma}) = \Phi\left(\frac{C - a_0}{\widehat{\sigma}}\right) = \alpha. \tag{2.5}$$

Since $\Phi\left(\frac{C-a_0}{\widehat{\sigma}}\right)$ is monotone increasing in C, condition (2.5) is equivalent to the rejection region

$$\mathcal{C} = \{\widehat{a} : \ \Phi\left(\frac{\widehat{a} - a_0}{\widehat{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{a})}\right) \leq \alpha\},$$

which was to prove.

3 Discrepancy measures and p-value-curves

3.1 Discrepancy measures

Before we are able to quantify the effects of overdispersion when a Poisson model is used, we have to find discrepancy measures, which can be used to compare the Poisson (POISSON)

with the negative binomial model (NB). This discrepancy measure should be chosen in such a way that its magnitude can be easily interpreted by the experimenter. The dispersion parameter a of the negative binomial distribution itself is however not suitable for this, since it is difficult to interpret its magnitude. We will denote this naive discrepancy measure with $d_N(a) = a$. Moreover, the discrepancy measure should be a monotone function of a to guarantee a unique interpretation. One possible measure is the ratio of the variances of both models:

$$\frac{Var(\text{NB})}{Var(\text{POISSON})} = \frac{\mu + a\mu^2}{\mu} = 1 + a\mu. \tag{3.1}$$

For regression data Y_i, $i = 1, .., n$ with means $\mu_i = \exp(\boldsymbol{x_i^t \beta})$, $i = 1, .., n$, we choose the maximum of (3.1) over μ_i as discrepancy measure given by

$$d(a) = 1 + a \max_i\{\mu_i\} = 1 + a \max_i\{e^{\boldsymbol{x_i^t \beta}}\}. \tag{3.2}$$

This discrepancy measure can be interpreted in the following way: If $d(a) = 2$, it follows that the maximal variance in the negative binomial regression model is twice as large as in the corresponding Poisson regression model. The experimenter can set a cut off value for $d(a)$ denoted by d_0. If $d(a) \leq d_0$, the experimenter will accept the Poisson regression model for the data, while if $d(a) > d_0$ a negative binomial regression will be accepted. The choice of reasonable values for d_0 will be discussed later.

Since the discrepancy measure $d(a)$ in (3.2) depends also on the unknown regression parameter $\boldsymbol{\beta}$, the value of $d(a)$ has to be estimated, one such possible estimate is given by

$$\widehat{d}(a) := 1 + a \max_i\{e^{\boldsymbol{x_i^t} \widehat{\boldsymbol{\beta}}(a)}\}, \tag{3.3}$$

where $\widehat{\boldsymbol{\beta}}(a)$ is the regression parameter estimate in the negative binomial regression model with fixed dispersion parameter a.

We now return to the problem of finding a reasonable cut off value d_0. For this we consider the consequences of fitting a false Poisson regression model to a data set which arises from a negative regression model with dispersion parameter a and approximately the same mean specification. In this case the regression coefficient estimates based on the incorrect Poisson model (denoted by $\widehat{\boldsymbol{\beta}}(0)$) and the correct negative binomial model (denoted by $\widehat{\boldsymbol{\beta}}(a)$) will be approximately the same, i.e. $\widehat{\boldsymbol{\beta}}(0) \approx \widehat{\boldsymbol{\beta}}(a)$. Using the asymptotic normality of the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}(a)$ in a negative binomial regression model with fixed dispersion a, it

follows that an estimate of the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}(a)$ denoted by $\widehat{cov}(\widehat{\boldsymbol{\beta}}(a))$ is given by

$$\widehat{cov}(\widehat{\boldsymbol{\beta}}(a)) = (X'D(a)X)^{-1}, \tag{3.4}$$

where $D(a)$ is a diagonal matrix with ith element given by

$$D_i(a) = \frac{\widehat{\mu_i}(a)}{1 + a\widehat{\mu_i}(a)}, \text{ where } \widehat{\mu_i}(a) = \exp(\boldsymbol{x_i^t}\widehat{\boldsymbol{\beta}}(a)).$$

Using the definition of $\widehat{d}(a)$ given in (3.3) and condition (3.4) it follows that

$$\widehat{cov}(\widehat{\boldsymbol{\beta}}(a)) \stackrel{\leq}{\approx} \widehat{d}(a)\widehat{cov}(\widehat{\boldsymbol{\beta}}(0)),$$

which means that the estimated asymptotic covariance matrix of the true regression parameter estimates $\widehat{\boldsymbol{\beta}}(a)$ is approximately underestimated by at most the factor $\widehat{d}(a)$ if the incorrect Poisson model is used. Since significance tests of covariate effects are based on such estimates of the standard error of the regression coefficients, it follows that a reasonable cut off point of d_0 might be around 2.

Before using the discrepancy measure defined in (3.2) in practice one has to check the composition of the data. Since this measure can take on very large values caused by a large maximal value for μ_i, it is useful to control the discrepancy measure for data with outliers in the mean space. In these cases we suggest to use instead of the maximal value $\max_i\{\mu_i\}$ of the mean some quantiles, like the 90% or 95% quantile of the means μ_i, yielding the following discrepancy measure:

$$d_q(a) = 1 + az_{q,\mu}, \tag{3.5}$$

where $z_{q,\mu}$ is the 100q%-th empirical quantile of $\{\mu_i, i = 1, .., n\}$ for $q \in (0, 1]$. Note that $d(a)$ is a special case of $d_q(a)$ with $q = 1$. Again this can be estimated by using $\widehat{\boldsymbol{\beta}}(a)$ to calculate the 100q%-th empirical quantile of $\{\widehat{\mu_i}(a) = e^{\boldsymbol{x_i^t}\widehat{\boldsymbol{\beta}}(a)}, i = 1, .., n\}$ as an estimate of $z_{q,\mu}$. Using (3.5) with the 90% quantile for example, one accepts, that the maximal deviation $d_q(a)$ in the variances of both models is valid only for 90% of the data.

3.2 p-value-curves

Even though we do not suggest to use $d_N(a) = a$ as discrepancy measure we first consider the testing problem

$$H : a > a_0 \quad \text{versus} \quad K : a \leq a_0 \quad \text{for fixed } a_0.$$

Using the asymptotic UMP-test derived in Theorem 2.1 we can calculate an estimate of the corresponding p-value, which is given by

$$\widehat{P}(a_0) = \Phi\left(\frac{\widehat{a} - a_0}{\widehat{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{a})}\right). \tag{3.6}$$

If we vary in the test problem H versus K the value of a_0, we can consider $\widehat{P}(.)$ as an asymptotic p-value curve. To interpret the p-value-curve $\widehat{P}(.)$, we consider the asymptotic p-value-curve for the testing problem $H : a > a_0$ versus $K : a \leq a_0$ and the testing problem $K : a \leq a_0$ versus $H : a > a_0$ respectively, in Figure 1. The testing problem H versus K

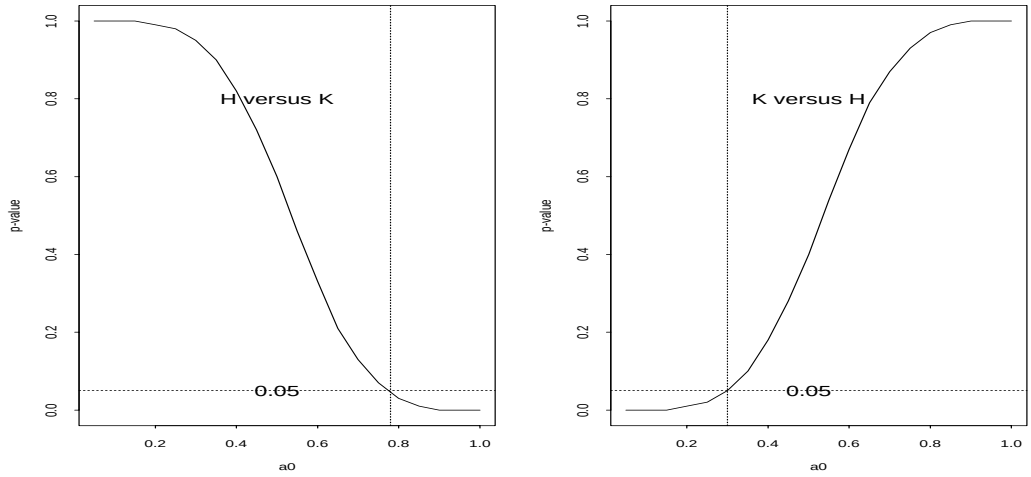

Figure 1: p-value-curves for the validation problem H versus K (left panel) and the discrimination problem K versus H (right panel)

can be considered as a **validation testing problem**, since if one rejects H, the dispersion parameter is validated to be less than a_0 with significance α. In a similar way, the testing problem K versus H can be regarded as a **discrimination testing problem**, since a rejection of K implies, that the dispersion parameter is larger than a_0 with significance α. In the left panel, the level $\widehat{P} = \alpha = 0.05$ determines the minimal value a_0, for which $H : a > a_0$ can be rejected at a level $\alpha = 0.05$ of significance. Looking at the right panel, it gives the maximal value a_0 at the level $\widehat{P} = \alpha = 0.05$ for rejecting $K : a \leq a_0$. Since the right curve can be generated by the left one by reflecting the curve at the axis $\widehat{P} = 0.5$, both test problems can be summarized in one curve (see Figure 2). Here, the value a_{0u} describes the minimal value for rejecting the hypothesis $H : a > a_0$ for a given level α of significance. At the same time it is also gives the maximal value a_{0l} at the level $\widehat{P} = 1 - \alpha$, for which the

hypothesis $K : a \leq a_0$ can be rejected for the given level α. The interpretation of Figure 2 can therefore be summarized as follows:

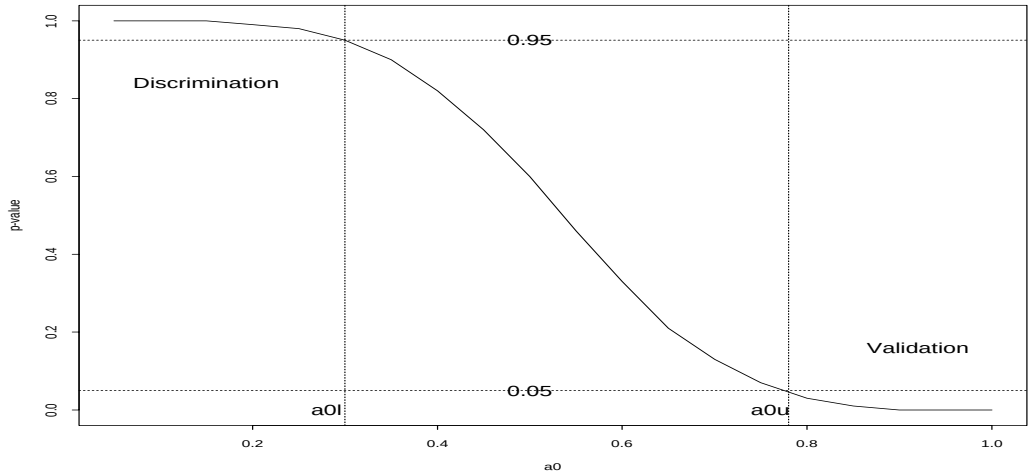

Figure 2: General p-value-curve

1. **Model validation ($H : a > a_0$ versus $K : a \leq a_0$):**
   The hypothesis $H : a > a_0$ can be rejected for values greater than a_{0u} at level α and the Poisson distribution can be used to model the given data if one is willing to accept a dispersion parameter less or equal a_{0u}.

2. **Model discrimination ($K : a \leq a_0$ versus $H : a > a_0$):**
   For values lower than a_{0l} the hypothesis $K : a \leq a_0$ can be rejected at level α and the change from a Poisson to a negative binomial model is justified if one assumes that a dispersion parameter $> a_{0l}$ indicates that a negative binomial is required.

In summary, the quantities a_{0u} and a_{0l} from the general p-value with respect to the dispersion parameter a allow to quantify the effect of overdispersion for count regression data when comparing Poisson regression to negative binomial regression. However, since the magnitude of the dispersion parameter is difficult to interpret we now propose to translate the concept of p-value curves to the more interpretable discrepancy measure $d(a)$. Similarly, the p-value-curve with respect to the testing problem $H : d(a) > d_0$ versus $K : d(a) \leq d_0$ is given by

$$\widehat{P}(d_0) = \Phi\left(\frac{\widehat{a} - d^{-1}(d_0)}{\widehat{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{a})}\right), \tag{3.7}$$

where $d^{-1}(.)$ is the inverse function of $d(.)$. If $d_q(a)$ is used in (3.7), d^{-1} is replaced by $d_q^{-1}(.)$, the inverse function of $d_q(.)$.

4 Examples

4.1 Automobile Claims

Hallin and Ingenbleek (1983) explored data on third party motor insurance claims in Sweden for the year 1977. Part of the data is reproduced in Andrews and Herzberg (1985) and the complete data is available from the Stalib database (`http://lib.stat.cmu.edu/`). The data were compiled by the Swedish Committee on the *Analysis of Risk Premium in Motor Insurance*. The dependent variable Y is the number of claims, the number of insured in policy years is used as offset (see Subsection 2.1) and possible risk variables are as follows:

- **kilometres**: kilometres travelled per year (categorized into 5 classes)
- **zone**: geographical zone (9 zones available)
- **bonus**: no claims bonus: equal to the number of years plus one, since last claim
- **make**: 1-8 represent eight different common car models. All other models are combined in class 9.

To compare the Poisson model with the negative binomial model we fit the data to both models. We use all given variables, since they are all significant. In this example we code all explanatory variables as categorical factors and we ignore interactions to prevent a high loss of degrees of freedom. For the Poisson regression model the residual deviance is given by 2966.1 on 2157 degrees of freedom. Since the residual deviance is considerably larger than the degree of freedom, overdispersion can be suspected. This is confirmed by a negative binomial regression model fit, which reduce the residual deviance to 2231.17 on 2156 degrees of freedom. However, the overdispersion parameter a is estimated to be $\hat{a} = 0.009$. Before one can conclude, that the lack of fit in the Poisson model is caused by overdispersion, one has to check the residuals and the specification of the link function. The plots of the deviance residuals of both models (not given here but given in Sikora 2002) are quite similar and clustered around zero, so that it can be excluded, that mean specification fits the data reasonably well. At last we checked the link specification with a plot of the number of claims against the linear predictor estimates, which also showed no lack of fit. Therefore,

it is reasonable to assume, that the high residual deviance is caused by overdispersion in the data. To quantify the overdispersion and to check, whether a Poisson regression model must be refused for this data, we determined the p-value curves, introduced in Section 3.2 given in Figure 3, corresponding to the modified rejection area given in (6.2). We modified the rejection area since the simulation results presented in the Appendix indicate that the original test is too liberal in small samples.

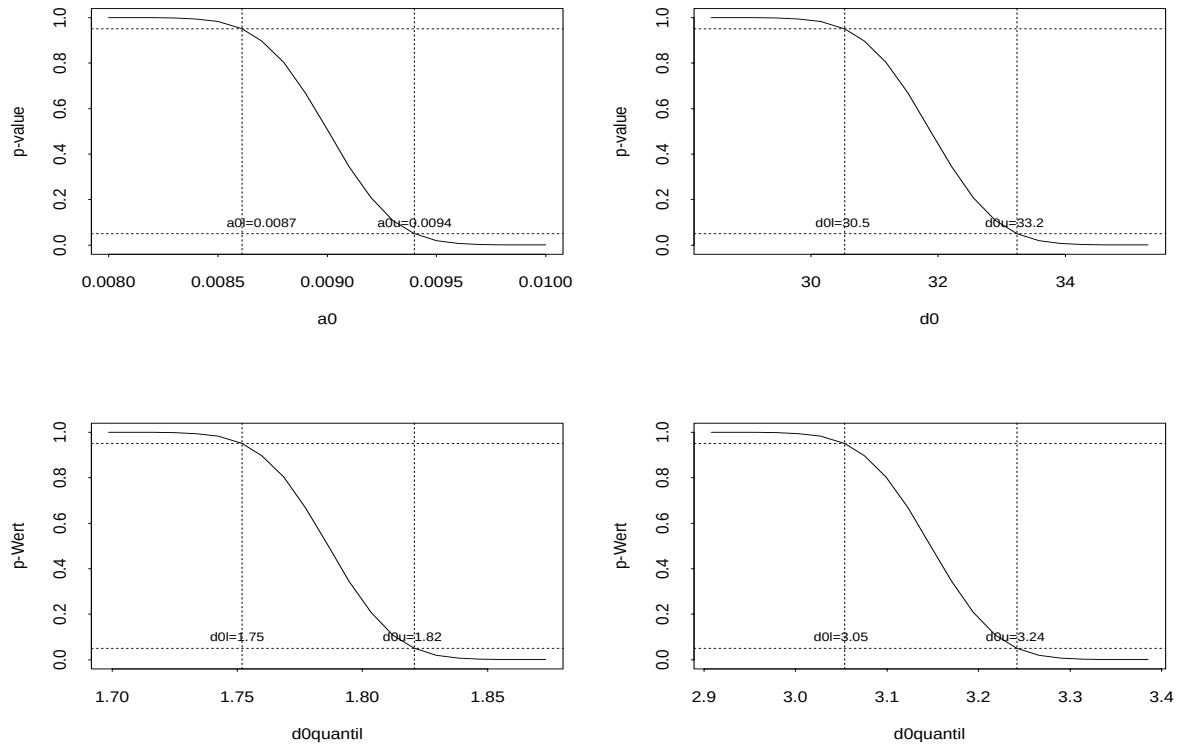

Figure 3: p-values using different discrepancy measures $d(.)$ (top left: $d_N(a)$, top right: $d(a)$, bottom left: $d_q(a), q = 0.90$, bottom right: $d_q(a), q = 0.95$) for the automobile claims data

The top left panel shows the p-value curve with respect to a. It illustrates, that one has to accept the minimal value $a_{0u} = 0.0094$ for rejecting the hypothesis $H : a > a_0$ to validate the Poisson model with a 10% significance level. Note the significance level is 10%, since we use (6.2) as rejection region. For this α level test the horizontal cut lines are are $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$. The value $a_{0u} = 0.0094$ seems to be very small, since $a = 0$ indicates a model without overdispersion, so that we would tend to accept the Poisson model. But we already noted, it is difficult to interpret the value a_{0u} and therefore it is more useful to consider the p-value curve respect to other discrepancy measures such as $d(a)$ or $d_q(a)$. The right panel in the

above line shows the curve respect to $d(a)$ with $d(a) = 1 + a \max_i\{\widehat{\mu_i}\}$. Here, it results in a value $d_0 = 33.2$ to reject the hypothesis $H : d > d_0$ at 10% significan level. Therefore, if one wants to validate the Poisson regression model, one has to accept, that the maximal variance of the negative binomial regression model is 33 times larger than the maximal variance of the Poisson model. This large value of d_0 can of course not be accepted, so that we are inclined to use negative binomial regression for this data. This example illustrates in an impressive way, how misleading the results based on the dispersion parameter estimate can be. It shows, how important it is, to use a discrepancy measure, which is easy to interpret. Finally, we check the composition of the data and find, that the large value of $d_0 = 33.2$ is caused by a large maximal mean estimate $\widehat{\mu}_{\max} = 3429.7$, whereas the 90% quantile of the estimated means $\widehat{\mu}_i$ is only 87 and 95% of the estimated means take on only values between 0 and 238. Therefore it is useful to consider the p-value curves respect to the discrepancy measure $d_q(a)$ as defined in (3.5). The bottom left panel shows, that for 90% of the data, the maximal deviation in the variances between the Poisson and the negative binomial regression model is only 1.82. In the bottom right panel the corresponding result for $q = 0.95$ are given, yielding a minimal value of $d_{0u} = 3.24$ here. This shows that if one is willing to disregard outlying observations causing a large expected mean value, then only a moderate increase in the variance needs to be tolerated when a Poisson regression model is used compared to a negative binomial regression model. If one does not want to disregard these observations, a change from the Poisson to the negative binomial model is required.

4.2 Patent data

Wang et al. (1998) investigated data on patents US high-tech firms in 1976. The dependent variable Y_i here is the number of patent applications and explanatory variables are R&D, which describes the annual research and development spending, and annual Sales. Using model selection techniques (for details see Sikora 2002) we select the following specification for $\mu_i = E(\mu_i)$

$$\log(\mu_i) = \beta_0 + \beta_1 R\&D/Sales + \beta_2 \sqrt[5]{R\&D} + \beta_3 R\&D/Sales * \sqrt[5]{R\&D} \qquad (4.1)$$

for our comparison. A Poisson regression model for mean specification (4.1) gives a residual deviance of 377.18 on 66 degrees of freedom, which is reduced to 78.55 on 65 degrees of freedom with an estimated overdispersion parameter $\widehat{a} = 0.29$. We checked as in the previous example that the better fit of the negative binomial model is likely to be caused by the fact

that it allows for overdispersion. While in the first example the dispersion parameter was estimated to be close to zero ($\widehat{a} = 0.009$), it is here estimated to be $\widehat{a} = 0.29$. We again determine the p-value curves corresponding to the rejection area (6.2) for this data set for different choices of discrepancy measures (Figure 4). The top left panel shows the p-

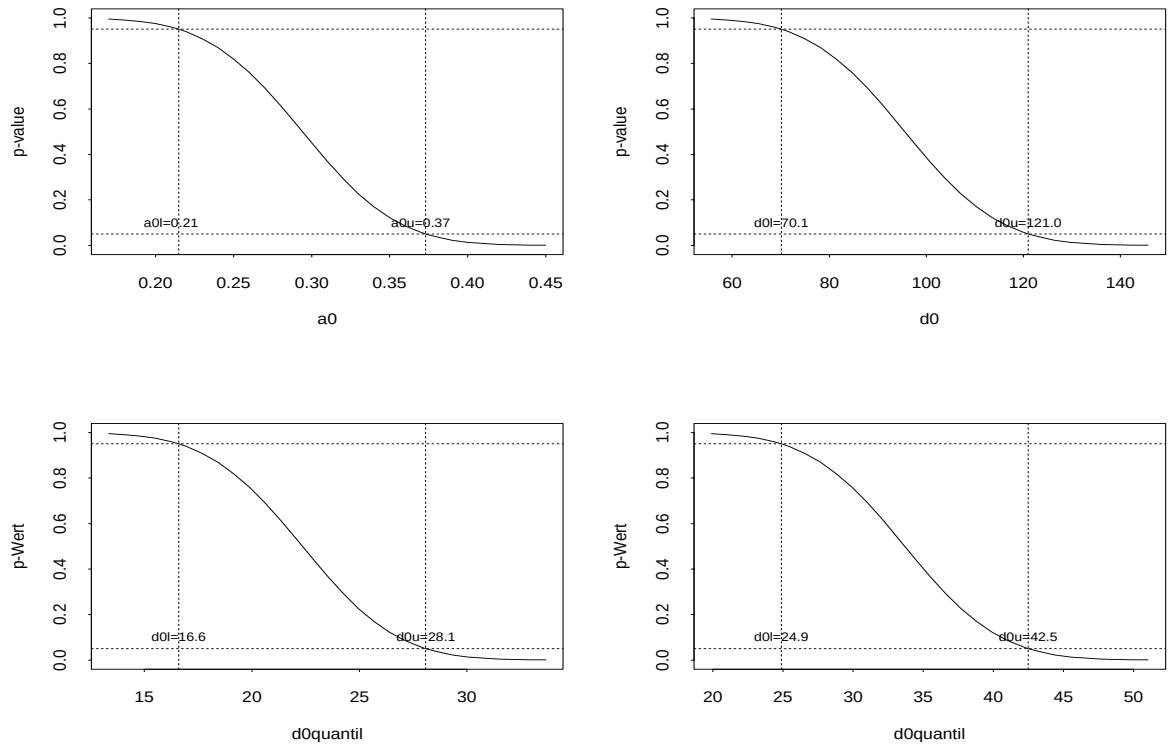

Figure 4: p-values using different discrepancy measures $d(.)$ (top left: $d_N(a)$, top right: $d(a)$, bottom left: $d_q(a), q = 0.90$, bottom right: $d_q(a), q = 0.95$) for the patent data

value curve respect to a. It results a minimal value $a_{0u} = 0.37$ for rejecting the hypothesis $H : a > a_0$ and to accept the Poisson model. In contrast to the first example (where the minimal value was $a_{0u} = 0.0094$) this value seems to be larger and we tend to reject the Poisson model. To quantify the overdispersion in a more interpretable way, we now consider the p-value curve respect to the discrepancy measure $d(a)$ given by the top right panel. As we have already expected, this figure confirms the rejection of the Poisson model, since the minimal value for rejecting the hypothesis $H : d > d_0$ at $\alpha = 10\%$ is $d_0 = 121$. This means that to validate the Poisson model, one has to accept, that the maximal variance in the negative binomial model is 121 times larger than the maximal variance in the Poisson model. This value is very large, so that we should use the negative binomial regression

to model this data. It remains to check the data for observations which produce extreme mean values. While the maximal estimated mean value is 321.7, 95% of the expected patent applications take on values between 0 and 111.2, and the 90%-quantile of the estimated means is 72.6. Since 90%- and 95%-quantiles of the estimated means are quite lower than the estimated maximum, it is useful to consider the p-value curves respect to the discrepancy measure $d_q(a)$. In the Figure 4 the bottom left panel shows the p-value curve with respect to the 90%-quantile, while the bottom right one gives the one with respect to the 95%-quantile. Both bottom panels result in large minimal values $d_{0u} = 28.1$ and $d_{0u} = 42.5$ respectively for rejecting the hypothesis $H : d > d_0$. This implies, that the maximal variances of both models differ by a factor of about 28 and 42, which can not be accepted. In contrast to the first example a consideration of the discrepancy measure $d_q(a)$ does not lead us to change our advise to use the negative binomial regression model instead of the Poisson regression model in this example. As we have already suspected from the large drop in the residual deviance, the quantification of the overdispersion effects by using p-value curves confirms in all panels of Figure 4 the rejection of the Poisson model and we advise to change to the more appropriate negative binomial regression model for this data set.

5 Summary and Discussion

We developed in this paper discrepancy measures between a Poisson regression model and a negative binomial regression model which can be easily interpreted by the experimenter. These measures can also be interpreted as giving approximate bounds on the underestimation of the standard errors of the regression parameters when a Poisson model is used while the true model is an negative binomial model. These measures are then used to construct asymptotic α-level tests for one sided null hypotheses, which allow for the quantification of overdispersion effects in terms of the interpretable discrepancy measures. Such a quantification of overdispersion effects cannot be achieved using score tests commonly developed for the detection of overdispersion.

In a simulation study (see Appendix) we show that the original proposed asymptotic test with rejection area (2.4) is too liberal in small sample sizes and as a result we modified the rejection area to (6.2), which gives approximately α-level tests in moderately large sample sizes ($n \geq 100$). We applied these tests to 2 data sets with moderately large sample sizes, where we feel comfortable to apply the modified test (6.2). Considering a modification of

the original discrepancy measure we can also adjust for the effects of outliers in the mean space.

Even though this paper concentrates on the quantification of overdispersion effects when a Poisson regression model is compared to a negative binomial regression model, the approach presented can be also used to consider other overdispersion models such as the generalized Poisson regression model (see Consul and Famoye 1992) or the double Poisson family of Efron 1986. Currently we investigate these extensions of the approach developed in the paper.

6 Appendix(*Simulation study*)

We investigated the small sample performance of the asymptotic test proposed in Theorem 2.1 through an extensive simulation study. To assess the small sample performance we determined estimates of the power function and the corresponding p-value curves for the discrepancy measures introduced in Section 3.1. In particular we examined the following questions:

- How does the sample size n influence the test performance?
- Does the range of means influence the test performance?
- Does a different value of the signal to noise ratio influence the results?

To answer these questions we generated negative binomial data sets $Y_i \sim NB(\mu_i, a)$ with $\mu_i = \exp(\boldsymbol{x_i^t \beta})$ and $\boldsymbol{x}_i^t = (1,\, x_i)$, i.e. only a single covariate was used. The values for $\boldsymbol{x_i}$, $i = 1, .., n$, were chosen equally spaced in a given interval. The data sets were generated with two different sample sizes, $n = 25$ and $n = 100$. For the range of the μ_i's we chose two different intervals. For the first range interval we chose the regression parameters β_0 and β_1 in such a way, that the resulting means μ_i vary in the interval $\frac{1}{2}\bar{\mu} < \mu_i < \frac{3}{2}\bar{\mu} \quad \forall\, i = 1, .., n$, where $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$. The second range for the means μ_i satisfies $\frac{1}{4}\bar{\mu} < \mu_i < \frac{7}{4}\bar{\mu} \quad \forall\, i = 1, .., n$. Finally, we considered the signal to noise ratio, which is defined as

$$SN(a, \mu_i) = \frac{E(Y_i)}{\sqrt{Var(Y_i)}}. \tag{6.1}$$

For the negative binomial case we have $SN_{NB}(a, \mu_i) = \frac{\mu_i}{\sqrt{\mu_i + a\mu_i^2}} = \sqrt{\frac{\mu_i}{1+a\mu_i}}$. We then fixed the two pairs $(a; \bar{\mu}) = (0.3; 3)$ and $(a; \bar{\mu}) = (0.63; 100)$, for which the resulting $SN(a, \bar{\mu})$ has a value of 1.25. With two other pairs $(a; \bar{\mu}) = (0.15; 10)$ and $(a; \bar{\mu}) = (0.24; 100)$ the

resulting value of $SN(a, \bar{\mu})$ is 2. Since the mean values are determined by the covariates and the regression parameters, the desired $SN(a, \bar{\mu})$ are only approximately achieved. The parameter sets $(a, \mu, n, \boldsymbol{x_i}, \boldsymbol{\beta})$, which are close to the described conditions are summarized in Table 1.

Simulation of the power function

At first we examined the behavior of the power function corresponding to the validation test for $H : a > a_0$ versus $K : a \leq a_0$ defined in Theorem 2.1. The power function is given by $\beta(a) = P(\text{reject } H : a > a_0)$ with the rejection region of the validation test given by $\Phi\left(\frac{\widehat{a}-a_0}{\widehat{\sigma}(\widehat{\boldsymbol{\beta}},\widehat{a})}\right) \leq \alpha$. For the hypothesis $H : a > a_0$ we chose 4 different values for a_0, namely $a_0 = 0.3, 0.63, 0.15$ and 0.24. For each parameter set in Table 1 we determined an estimate of the power function $\beta(a)$ based on 300 simulated data sets. The resulting 16 estimated power functions are summarized in 4 groups which have the same a_0. We carried out this study for levels of significance $\alpha = 0.05, 0.1$ and 0.15, but we present in Figure 5 only the results for $\alpha = 0.1$. The other results can be found in Sikora (2002, page 89-91).

At the vertical line $a = a_0$ one is accepting an error of $\alpha = 0.1$, i.e. the power function at this point should asymptotically not be larger than 0.1. Since all curves cut the vertical line above the value 0.1 (horizontal dotted line), it can be concluded, that the asymptotic test is liberal in small samples. Moreover, in all 4 groups it can be seen, that the sample size influences the test results, since the curves corresponding to a large sample size $n = 100$ (parameter sets 2,4,6,8,10,12,14 and 16) are steeper at $a = a_0$ and the corresponding tests are less liberal. To get information about the influence of the range of the means, one has to compare the 1st with the 3rd and the 2nd with the 4th parameter set in each group. The curves of the 1st and the 3rd parameter set, and of the 2nd and 4th respectively, are quite similar, so that we can conclude, that the range has minimal influence to the test results. Comparing the two panels in the first column with the two panels on the right, one can find out some differences in the level of errors at the vertical line $a = a_0$. The curves on the left side cut the line $a = a_0$ at a higher level than the curves on the right. The only difference in the parameter sets between these groups is the value of $\bar{\mu}$. While for the 1st and 3rd parameter set $\bar{\mu}$ was chosen as 3 and 10, it takes on the value 100 for the 2nd and 4th parameter set on the right. We conclude, that a larger $\bar{\mu}$ (see right side) induces a less

PS	$SN(a, \bar{\mu})$	a	$\bar{\mu}$	Range of μ_i	$d(a)$	n	Range of $\boldsymbol{x_i}$	(β_0, β_1)
1	1.25 desired	0.30	3.00	[1.5 ; 4.5]		25	[0.05, 1.25]	(0.3, 1)
	1.20 achieved		2.76	[1.42 ; 4.71]	2.41			
2	1.25 desired	0.30	3.00	[1.5 ; 4.5]		100	[-5, 4.9]	(1, 0.1)
	1.22 achieved		2.82	[1.65 ; 4.44]	2.33			
3	1.25 desired	0.30	3.00	[0.75 ; 5.25]		25	[-1.05, 2.6]	(0.5, 0.5)
	1.17 achieved		2.77	[0.98 ; 5.90]	2.77			
4	1.25 desired	0.30	3.00	[0.75 ; 5.25]		100	[-10.9, 9]	(1, 0.1)
	1.18 achieved		2.89	[0.91 ; 6.62]	2.99			
5	1.25 desired	0.63	100.00	[50 ; 150]		25	[15.2, 20]	(1, 0.2)
	1.25 achieved		95.70	[56.83 ; 148.41]	94.50			
6	1.25 desired	0.63	100.00	[50 ; 150]		100	[30.1, 40]	(1, 0.1)
	1.25 achieved		94.28	[55.15 ; 148.41]	94.50			
7	1.25 desired	0.63	100.00	[25 ; 175]		25	[11, 18.2]	(1.5, 0.2)
	1.25 achieved		91.09	[40.45 ; 170.72]	108.55			
8	1.25 desired	0.63	100.00	[25 ; 175]		100	[16, 30.85]	(2, 0.1)
	1.25 achieved		84.31	[36.60 ; 161.58]	102.79			
9	2.00 desired	0.15	10.00	[5 ; 15]		25	[-4.6, 5]	(2.2, 0.1)
	1.96 achieved		9.60	[5.70 ; 14.88]	3.23			
10	2.00 desired	0.15	10.00	[5 ; 15]		100	[-4.9, 5]	(2.2, 0.1)
	1.95 achieved		9.45	[5.53 ; 14.88]	3.23			
11	2.00 desired	0.15	10.00	[2.5 ; 17.5]		25	[-3.5, 3.7]	(2.2, 0.2)
	1.95 achieved		10.09	[4.48 ; 18.91]	3.84			
12	2.00 desired	0.15	10.00	[2.5 ; 17.5]		100	[-1.3, 13.6]	(1.5, 0.1)
	1.91 achieved		9.07	[3.94 ; 17.34]	3.61			
13	2.00 desired	0.24	100.00	[50 ; 150]		25	[1.7, 4.1]	(3, 0.5)
	1.99 achieved		91.30	[46.99 ; 156.02]	38.45			
14	2.00 desired	0.24	100.00	[50 ; 150]		100	[20.1, 30]	(2, 0.1)
	1.99 achieved		94.28	[55.15 ; 148.41]	36.62			
15	2.00 desired	0.24	100.00	[25 ; 175]		25	[39, 63]	(2, 0.05)
	2.00 achieved		100.90	[51.94 ; 172.43]	42.38			
16	2.00 desired	0.24	100.00	[25 ; 175]		100	[33.3, 63.2]	(2, 0.05)
	1.99 achieved		89.98	[39.06 ; 172.43]	42.38			

Table 1: Parameter sets (PS) used in the simulation study

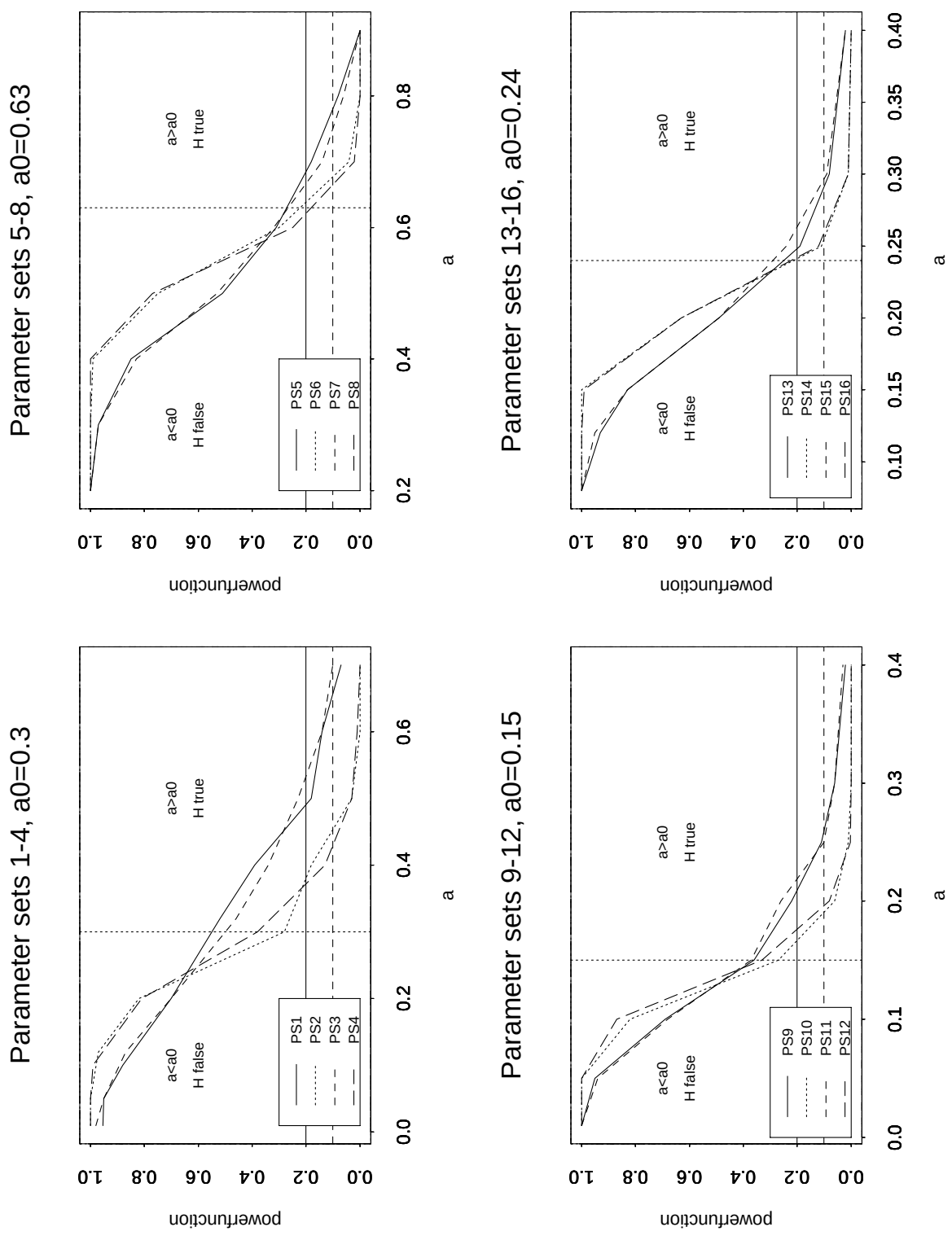

Figure 5: Estimated power functions for test (2.4) with $\alpha = 0.1$ (horizontal dotted line) and test (6.2) with $\alpha = 0.2$ (horizontal solid line)

liberal test. As we have seen in Figure 5, the validation test is a very liberal test in small samples. To reduce this problem, we propose to change the actual rejection region to the following rejection law:
Reject the hypothesis $H : a > a_0$ versus $K : a \leq a_0$ at a level of significance α, if and only if

$$\Phi\Big(\frac{\widehat{a} - a_0}{\widehat{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{a})}\Big) \leq \frac{\alpha}{2}. \tag{6.2}$$

With this change the test is less liberal, which can be seen at the horizontal solid line in Figure 5, which represents the new level of significance $\alpha = 0.2$. For the groups with a sample size $n = 100$ the curves cut the vertical line $a = a_0$ at a level around this value 0.2, so that we can accept this new test as an α-level test for larger sample sizes.

p-value curves respect to $d_N(a)$

Since we generate random variables the consideration of only a single p-value curve per parameter set can lead to wrong conclusions. Therefore, we repeat the construction of the p-value curves (3.6) 20 times for each parameter set and calculate the values $\bar{a}_{0u} = \frac{1}{20}\sum_{i=1}^{n} a_{0u_i}$ and $\bar{a}_{0l} = \frac{1}{20}\sum_{i=1}^{n} a_{0l_i}$ as the cut points of the curves with the levels $\widehat{P} = 0.05$ and $\widehat{P} = 0.95$. The values a_{0l_i} is the value for which in data set i the test based on (6.2) for $K : a > a_{0l_i}$ versus $H : a \leq a_{0l_i}$ is rejected at level $\alpha = .1$, while a_{0u_i} gives the value in data set i for which the test based on (6.2) rejects $H : a \leq a_{0u_i}$ versus $K : a > a_{0u_i}$ at $\alpha = .1$. Therefore $\bar{a}_{0l}$ gives the average value of $d_N(a)$ for which the negative binomial model can be discriminated from the Poisson model, while $\bar{a}_{0u}$ gives the average value of $d_N(a)$ for which the Poisson model can be validated. The values $\bar{a}_{0u}$, $\bar{a}_{0l}$ and the percentages of the deviations $a - \bar{a}_{0l}$ and $\bar{a}_{0u} - a$ are summarized in Table 2.
The p-value curves for all parameter sets can be found in Sikora (2002, page 97-100), while we present here only the curves of the first group with $a = 0.3$ and $\bar{\mu} = 3$. In Figure 6 the values $\bar{a}_{0l}$ and $\bar{a}_{0u}$ are shown by a vertical solid line and the true value of a is shown by a vertical dotted line.

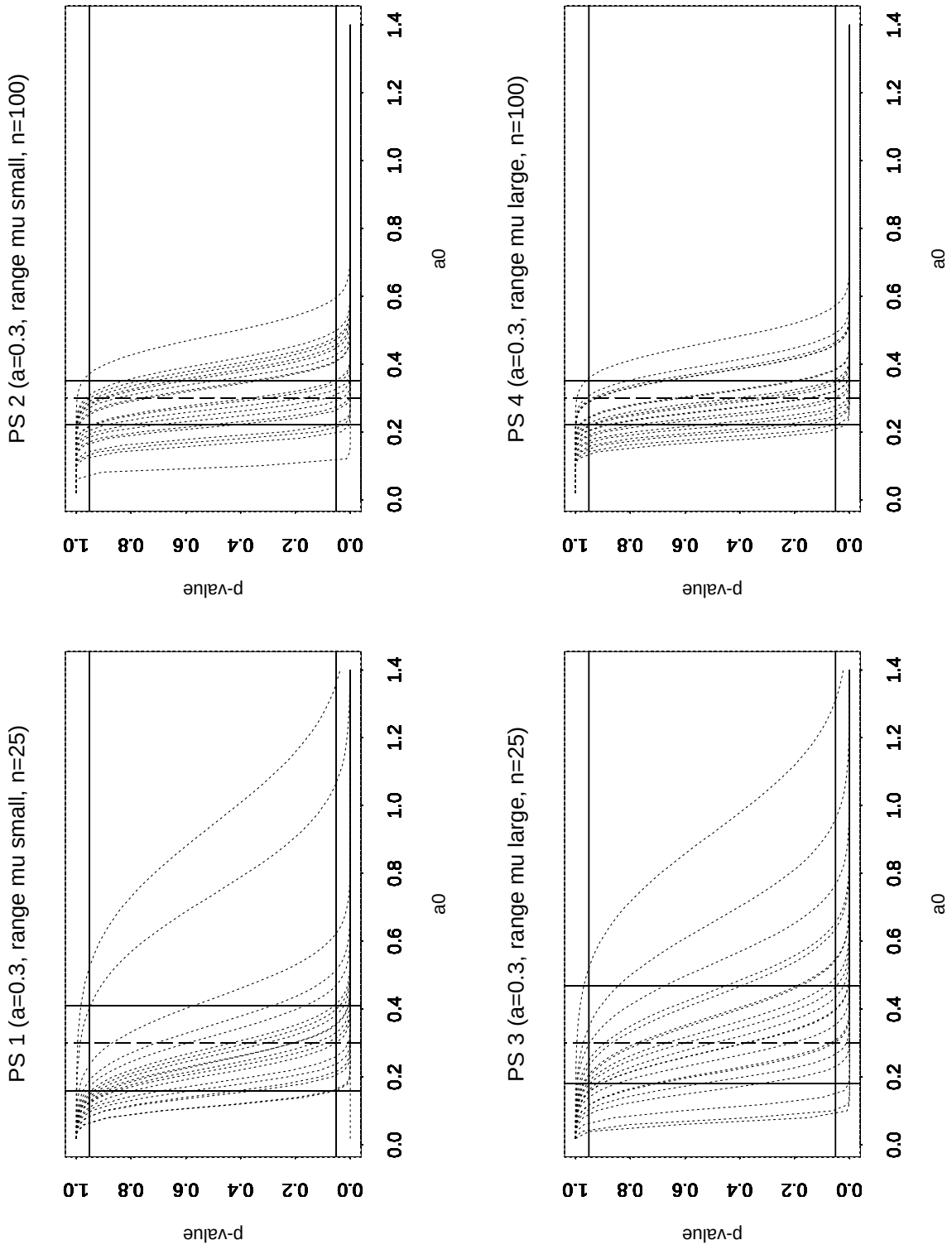

Figure 6: p-value curves with regard to $d_N(a) = a$ for the parameter sets (PS) 1-4

						Discrimination of Poisson		Validation of Poisson	
PS	n	$SN(a, \bar{\mu})$	$\bar{\mu}$	range interval μ_i	true a	$\bar{a}_{0l}$	true $a - \bar{a}_{0l}$ in %	$\bar{a}_{0u}$	$\bar{a}_{0u}$ − true a in %
1	25	1.25	3	[1.5 ; 4.5]	0.30	0.16	47%	0.41	37%
2	100	1.25	3	[1.5 ; 4.5]	0.30	0.22	27%	0.35	17%
3	25	1.25	3	[0.75 ; 5.25]	0.30	0.18	40%	0.47	57%
4	100	1.25	3	[0.75 ; 5.25]	0.30	0.22	27%	0.35	17%
5	25	1.25	100	[50 ; 150]	0.63	0.34	46%	0.88	40%
6	100	1.25	100	[50 ; 150]	0.63	0.47	25%	0.74	17%
7	25	1.25	100	[25 ; 175]	0.63	0.34	46%	0.89	41%
8	100	1.25	100	[25 ; 175]	0.63	0.49	22%	0.77	22%
9	25	2.00	10	[5 ; 15]	0.15	0.05	67%	0.16	7%
10	100	2.00	10	[5 ; 15]	0.15	0.11	27%	0.19	27%
11	25	2.00	10	[2.5 ; 17.5]	0.15	0.07	53%	0.21	40%
12	100	2.00	10	[2.5 ; 17.5]	0.15	0.09	40%	0.16	7%
13	25	2.00	100	[50 ; 150]	0.24	0.12	50%	0.33	38%
14	100	2.00	100	[50 ; 150]	0.24	0.18	25%	0.29	21%
15	25	2.00	100	[25 ; 175]	0.24	0.11	54%	0.31	29%
16	100	2.00	100	[25 ; 175]	0.24	0.17	29%	0.28	17%

Table 2: Results of the p-value curves respect to $d_N(a)$

The influence of the sample size can be seen by comparing the 1st with the 2nd and the 3rd with the 4th panel. In both cases the length of the interval $[\bar{a}_{0l}; \bar{a}_{0u}]$ is cut in half, switching from the curves of the parameter sets with sample size $n = 25$ (PS 1 and 3) to the curves of the parameter sets with $n = 100$ (PS 2 and 4). Therefore, the larger the sample size is, the smaller is the interval $[\bar{a}_{0l}; \bar{a}_{0u}]$. One can get the same results considering the values in Table 2, especially comparing the percentages of the deviations of $\bar{a}_{0l}$ and $\bar{a}_{0u}$ from a, respectively. Only the third group shows a somewhat different behavior, which might be explained by the randomness of the data. To decide which model should be used, the value $\bar{a}_{0u}$ should to be considered. In the 1st and 3rd parameter set with $n = 25$ a dispersion parameter index $a_0 = 0.41$ or 0.47 must be accepted to hold the Poisson model. But if one is willing to accept only a value $\bar{a}_{0l} = 0.16$ or 0.18, then the effort is justified to switch from the Poisson to the negative binomial model. For the other parameter sets with $n = 100$ the interval between $\bar{a}_{0l}$ and $\bar{a}_{0u}$ is much smaller, so that the area in which no decision can be taken, is reduced. In these cases one has to decide whether the value $\bar{a}_{0u}$ is low enough to accept and to validate the Poisson model. Comparing the first column of panels in Figure 6

with the second column one gets informations about the influence of the range of the means. Since the left panels are quite similar (the right panels respectively), we can conclude, that the range does not influence the shape of the p-value curves with respect to $d_N(a)$. Since it is difficult to interpret the magnitude of $d_N(a)$ we now present the results with regard to $d(a)$ which gives the maximal change in variances when changing from a Poisson model to a negative binomial model.

p-value curves respect to $d(a)$

As in Table 2 we summarized all results of the values $\bar{d}_{0u} = \frac{1}{20}\sum_{i=1}^{20} d_{0u_i}$ and $\bar{d}_{0l} = \frac{1}{20}\sum_{i=1}^{20} d_{0l_i}$ (cut points of the curves with levels $\widehat{P} = 0.05$ and $\widehat{P} = 0.95$), and their deviations $d - \bar{d}_{0l}$ and $\bar{d}_{0u} - d$ in percent in Table 3 for all parameter sets.

						Discrimination of Poisson		Validation of Poisson	
PS	n	$SN(a, \bar{\mu})$	$\bar{\mu}$	range interval μ_i	true d	d_{0l}	true $d - d_{0l}$ in %	d_{0u}	d_{0u} − true d in %
1	25	1.25	3	[1.5 ; 4.5]	2.41	1.63	32%	2.63	9%
2	100	1.25	3	[1.5 ; 4.5]	2.33	2.06	12%	2.69	15%
3	25	1.25	3	[0.75 ; 5.25]	2.77	2.19	21%	4.06	47%
4	100	1.25	3	[0.75 ; 5.25]	2.99	2.43	19%	3.25	9%
5	25	1.25	100	[50 ; 150]	94.50	52.31	45%	132.49	40%
6	100	1.25	100	[50 ; 150]	94.50	73.28	22%	114.06	21%
7	25	1.25	100	[25 ; 175]	108.55	56.17	48%	142.89	32%
8	100	1.25	100	[25 ; 175]	102.79	80.45	22%	125.31	22%
9	25	2.00	10	[5 ; 15]	3.23	1.84	43%	3.57	11%
10	100	2.00	10	[5 ; 15]	3.23	2.70	16%	3.89	20%
11	25	2.00	10	[2.5 ; 17.5]	3.84	2.22	42%	4.80	25%
12	100	2.00	10	[2.5 ; 17.5]	3.61	2.70	25%	3.82	6%
13	25	2.00	100	[50 ; 150]	38.45	19.80	49%	54.64	42%
14	100	2.00	100	[50 ; 150]	36.62	28.79	21%	46.14	26%
15	25	2.00	100	[25 ; 175]	42.38	18.99	55%	52.97	25%
16	100	2.00	100	[25 ; 175]	42.38	31.30	26%	50.49	19%

Table 3: Results of the p-value curves respect to $d(a)$

In Figure 3, which shows the first group of parameter sets, the true value of d is shown with a vertical dotted line, while the values $\bar{d}_{0l}$ and $\bar{d}_{0u}$ are shown with a vertical solid line. The horizontal solid lines represent the p-values $\widehat{P} = 0.05$ and $\widehat{P} = 0.95$. The p-value curves

of the other groups are given in Sikora (2002, page 107-110). Comparing again the 1st with the 2nd and 3rd with the 4th parameter set, we can see the influence of n, since the 1st and 3rd panels show the curves of parameter sets with $n = 25$ and the 2nd and 4th panels show curves of parameter sets with $n = 100$. In both cases the length of the interval $[\bar{d}_{0l}; \bar{d}_{0u}]$ is roughly cut in halves when switching from the 1st to the 2nd and from the 3rd to the 4th panel. Considering the percentages of deviations of $\bar{d}_{0l}$ and $\bar{d}_{0u}$ from d, we can find the same behavior as in Table 2. The larger n is, the smaller is the deviation. Only the deviation $\bar{d}_{0u} - d$ in Parameter Set 1 does not behave in this way, which can be seen in panel 1. In Group 3 (not shown here) we observe also some asymmetric behavior of the data. To decide, which model should be used, the value $\bar{d}_{0u}$ has to be considered. Comparing all 4 panels, the maximal value is $\bar{d}_{0u} = 4.06$ in Parameter Set 3. Accepting at most a deviation of 4 in the variances of both models, the Poisson model can be accepted in all 4 cases. This situation changes in Group 2 and 4 (see values $\bar{d}_{0u}$ in table 3). In these groups a Poisson model can not be accepted, since $\bar{d}_{0u}$ takes on values of 50 and more, which indicates a maximal variance of the negative binomial model 50 times larger than the maximal variance of the Poisson model.

The different results between the groups can be explained by the large $\bar{\mu} = 100$ in Group 2 and 4, which leads to very large values $\bar{d}_{0u}$, so that the Poisson model must be rejected. To get informations about the influence of the range of μ_i, we compare the 1st with 3rd and 2nd with 4th parameter set. The results in Table 3 confirm, that the larger the range of μ is (PS 2 and 4), the more the curves are moved to the right. The consequence of this behavior is, that in parameter sets with a large range of μ, the Poisson model is more difficult to validate, since $\bar{d}_{0u}$ takes on larger values.

Comparing now the first two groups with the last two groups to get results about the influence of the signal to noise ratio, we can not find any remarkable differences in the % deviations from the true discrepancy for similar $\bar{\mu}$ and same sample size n. We conclude that it is not the signal to noise ratio which influence the results of the test, but only the value of $\bar{\mu}$. The larger $\bar{\mu}$ is, the larger is $\bar{d}_{0u}$ and the more the Poisson model must be rejected in favor of the negative binomial model.

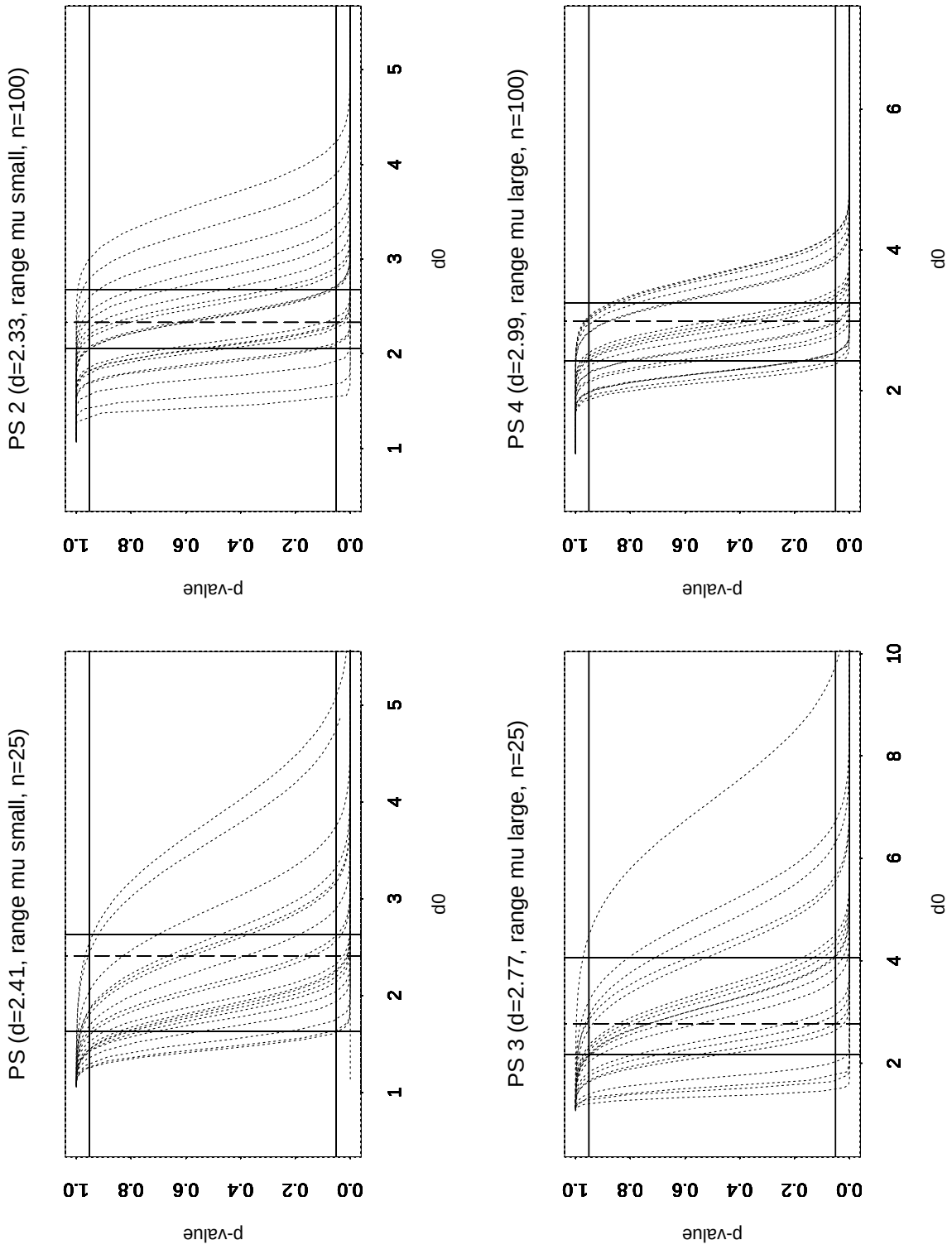

Figure 7: p-value curves with respect to $d(a)$ for the parameter sets (PS) 1-4

Summarizing the results of the simulation study, we see that the asymptotic test with rejection area (2.4) is too liberal in small samples, a modification of the rejection area to (6.2) gives a reasonable size $\alpha-$ level test in small samples with $n \geq 100$. A higher sample size produces steeper power functions and reduces the interval between discrimination and validation of the Poisson regression model when p-value curves are used. Finally, the signal to noise ratio has little influence both on power curves and p-value curves, while a larger $\bar{\mu}$ makes the tests based on (2.4) or (6.2) less liberal.

References

Andrews, D. F. and A. M. Herzberg (1985). *Data. A collection of problems from many fields for the student and reseach worker*. Springer.

Aragón, J., D. Eberly, and S. Eberly (1992). Existence and uniqueness of the maximum likelihood estimator for the two-parameter negative binomial distribution. *Statistics & Probability Letters 15*, 375–379.

Armitage, P. and T. Colton (1998). *Encyclopedia of Biostatistics, Vol.4*. John Wiley & Sons.

Besag, J., J. York, and A. Mollie (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics 43*, 1–59.

Breslow, N. and D. Clayton (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc. 88*, 9–25.

Cameron, A. and P. Trivedi (1998). *Regression analysis of count data*. Cambridge University Press.

Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *J. Appl. Econometrics 1*, 29–53.

Chesher, A. (1984). Testing for neglected heterogeneity,. *Econometrica 52*, 865–872.

Chow, S.-C. and J.-P. Liu (1992). *Design and analysis of bioavailability and bioequivalence studies*. Marcel Dekker.

Clark, S. J. and J. N. Perry (1989). Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics 45*, 309–316.

Clayton, D. G. (1996). Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice (eds. Gilks, W.R., Richardson, S. and Spiegelhalter, D.J.)*, pp. 275–301. Chapman & Hall.

Collings, B. J. and B. H. Margolin (1985). Testing goodness of fit for the poisson assumption when observations are not identically distributed. *J. Amer. Statist. Assoc. 80*, 411–418.

Consul, P. C. (1989). *Generalized Poisson distributions: Properties and applications*. Marcel Dekker.

Consul, P. C. and F. Famoye (1992). Generalized poisson regression model. *Communication in Statistics Series A 21*, 89–109.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika 70*, 643–649.

Czado, C. and A. Munk (2000). Noncanonical links in generalized linear models - when is the effort justified. *J. Statist. Plann. Inference 87*, 317–345.

Dean, C. (1992). Testing for overdispersion in poisson and binomial regression models. *J. Amer. Statist. Assoc. 87*, 451–457.

Dean, C. and J. Lawless (1989). Tests for detecting overdispersion in poisson regression models. *J. Amer. Statist. Assoc. 84*, 467–472.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc. 81*, 709–721.

Fisher, R. (1950). The significance of deviations from expectation in a poisson series. *Biometrics 6*, 17–24.

Fitzmaurice, G. M. (1997). Model selection with overdispersed data. *Statistician 46*, 81–91.

Ganio, L. and D. Schafer (1992). Diagnostics for overdispersion. *J. Amer. Statist. Assoc. 87*, 795–804.

Greenwood, M. and G. Yule (1920). An enquiry into the nature of frequency distributions and multiple happenings with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *J. Roy. Statist. Soc. A 83*, 255–279.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive poisson regression model. *J. Bus. Econom. Statist. 9*, 215–222.

Gurmu, S. and P. Trivedi (1992). Overdispersion test for truncated poisson regression models. *J. Econometrics 54*, 347–370.

Hallin, M. and J.-F. Ingenbleek (1983). The swedish automobile portfolio in 1977. *Scandinavian Actuarial Journal*, 49–64.

Johnson, N., S. Kotz, and A. Kemp (1993). *Univariate discrete distributions*. John Wiley & Sons.

Lambert, D. and K. Reoder (1995). Overdispersion diagnostics for generalized linear models. *J. Amer. Statist. Assoc. 90*, 1225–1236.

Lawless, J. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics 15*, 209–225.

Lehmann, E. (1986). *Testing statistical hypotheses*. John Wiley & Sons.

Lindsay, B. G. (1986). Exponential family mixture models (with least-squares estimators). *Annals of Statistics 14*, 124–137.

McCullagh, P. and J. Nelder (1989). *Generalized linear models*. Chapman and Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc. 92*, 162–170.

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

Sikora, I. (2002). *Quantifizierung von Überdispersion*. Munich: Master thesis at the University of Munich, `(http://www-m4.mathematik.tu-muenchen.de/m4/Diplarb/)`.

Singh, K. P. and F. Famoye (1993). Analysis of rates using a generalized poisson regression model. *Biometrical J. 35*, 917–923.

Wang, P., I. Cockburn, and M. Puterman (1998). Analysis of patent data - a mixed-poisson-regression-model approach. *J. Bus. Econom. Statist. 16*, 27–41.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects: A gibbs sampling approach. *J. Amer. Statist. Assoc. 86*, 79–86.