Smooth Tests of Fit for a Mixture of Two Poisson Distributions

D.J. Best, J.C.W. Rayner
The University of Newcastle, Callaghan, NSW, 2308, Australia
John.Best@newcastle.edu.au and John.Rayner@newcastle.edu.au

and O.Thas
Department of Applied Mathematics, Biometrics and Process Control, B-9000 Gent, Belgium
Olivier.Thas@Ugent.be

Abstract

In this note smooth tests of fit for a mixture of two Poisson distributions are derived and compared with a traditional Pearson chi-squared test. The tests are illustrated with a classic data set of deaths per day of women over 80 as recorded in the London Times for the years 1910 to 1912.

Keywords: Central moments, Deaths of London women data, Factorial moments, Orthonormal polynomials, Parametric bootstrap, Pearson X^2 tests

1. Introduction

A Poisson process is often used to model count data. Sometimes an underlying mechanism suggests two Poisson processes may be involved. This may be modelled by a two component Poisson mixture model. The two Poisson mixture applies generally to data more dispersed than that modelled by a single Poisson. An interesting example was given by Leroux and Puterman (1992) who fit a two Poisson mixture to fetal lamb movement data. They say the mixture model "... has a clear interpretation in terms of a ... background rate ... and an excited state." The Poisson probability function, $f(x;\ \theta)$ say, is given by

$$f(x;\ \theta) = \exp(-\theta)\theta^x/x!,\ x = 0, 1, 2, ...,$$
$$\text{in which } \theta > 0$$

and the two component Poisson mixture model has probability function

$$f^*(x;\ \theta_1,\ \theta_2,\ p) = p\ f(x;\ \theta_1) + (1 - p)\ f(x;\ \theta_2),$$
$$x = 0, 1, 2, ..., \text{ in which } \theta_1 > 0,\ \theta_2 > 0,$$
$$\theta_1 \neq \theta_2 \text{ and } 0 < p < 1.$$

A common test of fit for $f^*(x;\ \theta_1,\ \theta_2,\ p)$ is based on the well-known Pearson's X^2. If there are l classes X^2 is approximately χ^2 with $l - 4$ degrees of freedom: χ^2_{l-4}.

In section 2 we look at estimation of the parameters θ_1, θ_2 and p. Section 2 also defines the X^2 test and some smooth tests of fit. Section 3 gives a small power comparison while section 4 considers a classic data set of deaths per day of women over 80 as recorded in the London Times for the years 1910 to 1912.

2. Estimation and Test Statistics

The two most common approaches for estimating θ_1, θ_2 and p are based on moments (MOM) and maximum likelihood (ML). If we have n data points $x_1,\ x_2,\ \ldots,\ x_n$ and $\bar{x} = \sum_{i=1}^{n} x_i / n$ and $m_t = \sum_{i=1}^{n} (x_i - \bar{x})^t / n$, $t = 2, 3, ...$ the MOM estimators satisfy

$$\tilde{p} = (\bar{x} - \tilde{\theta}_2)/(\tilde{\theta}_1 - \tilde{\theta}_2),\ \tilde{\theta}_1 = (A - D)/2,$$
$$\text{and } \tilde{\theta}_2 = (A + D)/2$$

in which

$$A = 2\bar{x} + (m_3 - 3m_2 + 2\bar{x})/(m_2 - \bar{x})$$
$$\text{and } D^2 = A^2 - 4A\bar{x} + 4(m_2 + \bar{x}^2 - \bar{x}).$$

This method clearly fails if $D^2 < 0$, if any of $\tilde{\theta}_1$, $\tilde{\theta}_2$ and $\tilde{p}$ are outside their specified bounds, or if $m_2 = \bar{x}$.

Iteration is needed to find the ML estimates and given the speed of modern computers an EM type algorithm is satisfactory. This will always converge to $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{p}$ within the specified bounds if the initial estimates are also within these bounds. However convergence can be slow – occasionally more than 1,000 iterations – and a local, but not universal, maximum may be found. A grid of initial values is often worth examining. This was not done for the calculations in Table 1 because all the sizes were 0.05 suggesting universal maxima were indeed found. To check on the possibility of a local stationary point it is also useful to examine contour plots of the likelihood surface. This was done for the *Deaths of London Women* example in section 4. The following estimation equations are needed:

$$\hat{p}_k = \frac{\hat{p}_{k-1}\sum_{i=1}^{n} f(x_i, \hat{\theta}_{1,k-1})}{nf^*(x_i; \hat{\theta}_{1,k-1}, \hat{\theta}_{2,k-1}, \hat{p}_{k-1})} \text{ and}$$

$$\hat{\theta}_{r,k} = \frac{\sum_{i=1}^{n} x_i f(x_i, \hat{\theta}_{r,k-1})}{nf^*(x_i; \hat{\theta}_{1,k-1}, \hat{\theta}_{2,k-1}, \hat{p}_{k-1})}, \; r = 1, 2,$$

where $\hat{\theta}_{r,k}$ is the estimate of θ_r at the kth iteration, $\hat{p}_k$ is the estimate of p at the kth iteration and $(\hat{\theta}_{1,0}, \hat{\theta}_{2,0}, \hat{p}_0) = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{p})$ may be an admissible initial value. See, for example, Everitt and Hand (1981, p.97). Newton's method will sometimes converge to the correct values and when it does the convergence is much quicker than the above estimating equations. However, Newton's method doesn't always converge and may give estimates outside the specified bounds.

Now let O_j be the number of data points equal to j, $j = 0, 1, 2, \ldots$. Let $E_j = nf^*(j; \hat{\theta}_1, \hat{\theta}_2, \hat{p})$. Often classes are pooled in the tail until the greatest l is found such that the expectation of the classes from the lth on is at least 5. Then the Pearson test of fit statistic is

$$X^2 = \sum_{j=1}^{l} (O_j - E_j)^2 / E_j$$

and X^2 is taken to have the χ^2_{l-4} distribution.

Smooth test components V_s can be defined as

$$\hat{V}_s = \sum_{i=1}^{n} g_s(x_i; \hat{\theta}_1, \hat{\theta}_2, \hat{p}) / \sqrt{n}, \; s = 2, 3, \ldots$$

Here $\{g_s(.)\}$ is the set of orthonormal functions on the null distribution. We give formulae, in terms of the population moments $\mu, \mu_2, \ldots, \mu_6$ for the first four orthonormal functions and $\hat{V}_2$ and $\hat{V}_3$ in Appendix A. For the mixture of two Poissons these moments can be calculated from the population factorial moments $\mu'_{[t]} = p\theta_1^t + (1-p)\theta_2^t$. Smooth tests of fit are discussed in detail in Rayner et al. (2009).

Table 1. 100×powers based on 10,000 Monte Carlo samples for $n = 100$ and $\alpha = 0.05$ for a null Poisson mixture with $p = 0.5$, $\theta_1 = 2$ and $\theta_2 = 5$.

Alternative	$\hat{V}_2^2$	$\hat{V}_3^2$	$\hat{V}_4^2$	X^2
Null	5	5	5	5
NB(2, 0.4)	45	39	40	41
NB(3, 0.5)	18	20	20	18
NB(4, 0.5)	19	20	24	27
NTA(1, 2)	79	69	51	54
0.5 × NB(2, 0.4) + 0.5 × NB(2, 0.5)	33	28	30	31
0.5 × NB(2, 0.3) + 0.5 × NB(3, 0.5)	64	48	59	65
NTA(2, 2)	88	66	55	81
NTA(2, 1)	26	26	22	16
NTA(1, 3)	98	94	72	92
P(4)	37	14	13	4
P(6)	33	13	5	10

3. Indicative Size and Power Study

We consider the case $\alpha = 0.05$, $p = 0.5$, $\theta_1 = 2$, $\theta_2 = 5$. Based on 25,000 Monte Carlo samples the critical values of $\hat{V}_2^2$, $\hat{V}_3^2$ and $\hat{V}_4^2$ are 0.31, 0.91 and 0.56 respectively. We note that $\hat{V}_1 \equiv 0$ as shown in Appendix B. We use 17.5 as the X^2 critical value. Table 1 gives some powers for

- negative binomial alternatives with probability function $\binom{m+x-1}{x}\pi^m(1-\pi)^x$ for x = 0, 1, 2, ... with $m > 0$, denoted by NB(m, π),
- Neyman Type A alternatives with probability function $\frac{e^{-\lambda_1}\lambda_2^x}{x!}\sum_{j=0}^{\infty}\frac{j^x}{j!}(\lambda_1 e^{-\lambda_2})^j$ for x = 0, 1, 2, ... with $\lambda_1 > 0$ and $\lambda_2 > 0$, denoted by NTA(λ_1, λ_2) and
- Poisson alternatives $f(x; \theta)$ for x = 0, 1, 2, ... with $\theta > 0$, denoted by P(θ).

In Table 1 no one test dominates but overall perhaps that based on $\hat{V}_2^2$ does best. Double precision arithmetic was used in the Table 1 calculations. In a few cases no estimate was obtained after 10,000 iterations and these cases were discarded.

4. Example: Deaths of London Women During 1910 to 1912

A classic data set considered by a number of authors starting with Whitaker (1914) considers deaths per day of women over 80 in London during the years 1910, 1911 and 1912 as recorded in the Times newspaper. Table 2 shows the data and expected counts for $(\hat{\theta}_1, \hat{\theta}_2, \hat{p})$ = (1.257, 2.664, 0.360). Using ten classes X^2 = 1.29 with six degrees of freedom and χ^2 p-value 0.65. Also $\hat{V}_2^2$ = $(-0.077)^2$, $\hat{V}_3^2$ = $(-0.314)^2$ and $\hat{V}_4^2$ = $(-0.429)^2$, with bootstrap p-values 0.70, 0.46 and 0.55 respectively. Possibly due to different death rates in summer and winter, all tests indicate a good fit by a Poisson mixture. If a single Poisson is used to describe the data then X^2 = 27.01 with eight degrees of freedom and a χ^2 p-value is 0.001.

Table 2. Deaths per day of London women over 80 during 1910 to 1912

# deaths	0	1	2	3	4
Count	162	267	271	185	111
Mixture expected	161	271	262	191	114
Poisson expected	127	273	295	212	114

# deaths	5	6	7	8	9
Count	61	27	8	3	1
Mixture expected	58	25	9	3	1
Poisson expected	49	18	5	1	0

A plot of likelihood contours indicated the likelihood has a maximum at $(\hat{\theta}_1, \hat{\theta}_2)$ and that there are no other stationary points nearby. As $\hat{V}_1 \equiv 0$ we can give $\hat{p}$ in terms of $\hat{\theta}_1$ and $\hat{\theta}_2$ and so $\hat{p}$ does not need to be included in any likelihood contour plot.

References

[1] Everitt, B.S. and Hand, D.J.C. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

[2] Leroux, B.G. and Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, 48(2), 545-558.

[3] Rayner, J.C.W., Thas, O. and Best, D.J. (2009). *Smooth Tests of Goodness of Fit: Using R* (2nd ed.). Singapore: Wiley.

[4] Whitaker, L. (1914). On the Poisson law of small numbers. *Biometrika*, 10(1), 36-71.

Appendix A: Orthonormal Polynomials for a Poisson Mixture

Let μ be the mean and μ_t for t = 2, 3, ... the central moments, assumed to exist, of some distribution of interest. Then the first four orthonormal polynomials are, for x = 0, 1, 2, ...

$$g_0(x) = 1,\ g_1(x) = (x-\mu)/\sqrt{\mu_2},$$
$$g_2(x) = \{(x-\mu)^2 - \mu_3(x-\mu)/\mu_2 - \mu_2\}/\sqrt{d}$$
$$\text{and } g_3(x) = \{(x-\mu)^3 - a(x-\mu)^2 - b(x-\mu) - c\}/\sqrt{e}$$

where

$$d = \mu_4 - \mu_3^2/\mu_2 - \mu_2^2 \text{ and } e = \mu_6 - 2a\mu_5 + (a^2 - 2b)\mu_4 + 2(ab - c)\mu_3 + (b^2 + 2ac)\mu_2 + c^2$$

in which

$$a = (\mu_5 - \mu_3\mu_4/\mu_2 - \mu_2\mu_3)/d$$
$$b = (\mu_4^2/\mu_2 - \mu_2\mu_4 - \mu_3\mu_5/\mu_2 + \mu_3^2)/d$$
$$c = (2\mu_3\mu_4 - \mu_3^3/\mu_2 - \mu_2\mu_5)/d.$$

Again assuming they exist, for t = 2, 3, ... write $\mu'_{[t]}$ for the tth factorial moment. It now follows routinely that

$$\mu_2 = \mu'_{[2]} + \mu - \mu^2$$

$$\mu_3 = \mu'_{[3]} + 3\mu'_{[2]} + \mu - 3\mu(\mu'_{[2]} + \mu)\,\mu_2 + 2\mu^3$$

$$\mu_4 = \mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu - 4\mu(\mu'_{[3]} + 3\mu'_{[2]} + \mu) + 6\mu^2(\mu'_{[2]} + \mu) - 3\mu^4$$

$$\mu_5 = \mu'_{[5]} + 10\mu'_{[4]} + 25\mu'_{[3]} + 15\mu'_{[2]} + \mu - 5\mu(\mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu) + 10\mu^2(\mu'_{[3]} + 3\mu'_{[2]} + \mu) - 10\mu^3(\mu'_{[2]} + \mu) + 4\mu^5$$

$$\mu_6 = \mu'_{[6]} + 15\mu'_{[5]} + 65\mu'_{[4]} + 90\mu'_{[3]} + 31\mu'_{[2]} + \mu - 6\mu(\mu'_{[5]} + 10\mu'_{[4]} + 25\mu'_{[3]} + 15\mu'_{[2]} + \mu) + 15\mu^2(\mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu) - 20\mu^3(\mu'_{[3]} + 3\mu'_{[2]} + \mu) + 15\mu^4(\mu'_{[2]} + \mu) - 5\mu^6.$$

For a Poisson mixture the tth factorial moment is $\mu'_{[t]} = p\theta_1^t + (1-p)\theta_2^t$ so that, for example, $\mu = p\theta_1 + (1-p)\theta_2$. Using the ML estimators $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{p}$ and the above formulae for μ, ..., μ_6 we can calculate $\hat{V}_2$ and $\hat{V}_3$ where $\hat{V}_s = \sum_{i=1}^n g_s(x_i;\hat{\theta}_1,\hat{\theta}_2,\hat{p})/\sqrt{n}$, s = 2, 3.

Appendix B: Proof That $\hat{V}_1 \equiv 0$

Given $g_1(x)$ from Appendix A above, the first smooth component $\sum_{i=1}^n g_1(x_i;\hat{\theta}_1,\hat{\theta}_2,\hat{p})/\sqrt{n}$ is proportional to $\bar{X} - \hat{\mu}$, where $\hat{\mu} = \hat{p}\hat{\theta}_1 + (1-\hat{p})\hat{\theta}_2$ is the ML estimator of μ = E[X]. For notational convenience arguments involving θ_1, θ_2 and p are henceforth suppressed. To obtain the ML estimators of θ_1, θ_2 and p note that the likelihood is $L = \prod_{i=1}^n f^*(x_i;\theta_1,\theta_2,p)$. Taking logarithms and differentiating gives

$$\frac{\partial \log L}{\partial \theta_1} = \sum_{i=1}^n \{pf_1(x_i)[-1+\frac{x_i}{\theta_1}]\}/f^*(x_i),$$

$$\frac{\partial \log L}{\partial \theta_2} = \sum_{i=1}^n \{pf_2(x_i)[-1+\frac{x_i}{\theta_2}]\}/f^*(x_i)$$

and

$$\frac{\partial \log L}{\partial p} = \sum_{i=1}^n \{f_1(x_i) - f_2(x_i)/f^*(x_i).$$

From $\partial \log L/\partial\theta_r = 0$ for r = 1 and 2 we obtain

$$\hat{\theta}_r = \frac{\sum_{i=1}^n x_i f_r(x_i)/f^*(x_i)}{\sum_{i=1}^n f_r(x_i)/f^*(x_i)}$$

and from $\partial \log L/\partial p = 0$ we obtain

$$\sum_{i=1}^n f_1(x_i)/f^*(x_i) = \sum_{i=1}^n f_2(x_i)/f^*(x_i).$$

Using $f^*(x) = p\,f_1(x) + (1-p)\,f_2(x)$ and the equation immediately above shows that $\sum_{i=1}^n f_r(x_i)/f^*(x_i) = n$ for r = 1 and 2. It now follows that

$$\hat{\theta}_r = \sum_{i=1}^n x_i f_r(x_i)/\{nf^*(x_i)\} \text{ and}$$

$$\hat{\mu} = \hat{p}\hat{\theta}_1 + (1-\hat{p})\hat{\theta}_2 =$$

$$\hat{p}\sum_{i=1}^n x_i f_1(x_i)/f^*(x_i)/n +$$

$$(1-\hat{p})\sum_{i=1}^n x_i f_2(x_i)/f^*(x_i)/n =$$

$$\sum_{i=1}^n x_i\{\hat{p}f_1(x_i) + (1-\hat{p})f_2(x_i)\}/\{nf^*(x_i)\} =$$

$$\sum_{i=1}^n x_i/n = \bar{x}.$$

It thus follows that $\hat{V}_1 \equiv 0$.

Assessing Poisson and Logistic Regression Models Using Smooth Tests

Paul Rippon, J.C.W. Rayner

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
paul.rippon@newcastle.edu.au

Abstract

A smooth testing approach has been used to develop a test of the distributional assumption for generalized linear models. Application of this test to help assess Poisson and logistic regression models is discussed in this paper and power is compared to some common tests.

Key words: generalized linear models, goodness of fit, logistic regression, Poisson regression

1. Introduction

The concept of smooth testing originally proposed in [1] has been developed in [2] to provide goodness of fit tests for a wide range of distributions. In [3], these ideas have been applied to the generalized linear modelling framework, where the variables are no longer identically distributed, to derive a test of the distributional assumption. Section 2 describes the test, Section 3 comments on its application and Section 4 discusses the results of simulation studies examining the power of this test when applied to Poisson and logistic regression.

2. A Smooth Test of the Distributional Assumption in Generalized Linear Models

The generalized linear modelling structure comprises a linear combination of predictor variables related via a link function to the mean of the response distribution selected from the exponential family of distributions. In commonly used notation, independent response variables, $Y_1, \ldots, Y_n$, are distributed with density function

$$f(y_j;\theta_j) = \exp\left[\frac{y_j\theta_j - b(\theta_j)}{a(\phi_j)} + c(y_j,\phi_j)\right]$$

from an exponential family with canonical parameters θ_j to be estimated and dispersion parameters ϕ_j assumed to be known; a, b and c are known functions. Using $g(\cdot)$ to represent the link function:

$$g(\mu_j) = \eta_j = \boldsymbol{x}_j^T\boldsymbol{\beta} = x_{j1}\beta_1 + \ldots + x_{jp}\beta_p$$

where $\mu_j = E[Y_j] = b'(\theta_j)$ for $j = 1,\ldots,n$. To simplify subscripting, an explicit intercept term, β_0, is not shown. There is no loss of generality as β_1 can become an intercept term by setting all $x_{j1} = 1$.

To test the distributional assumption, the assumed response variable density, $f(y_j;\theta_j)$, is embedded within a more complex alternative density function

$$f_k(y_j;\boldsymbol{\tau},\theta_j) = C(\boldsymbol{\tau},\theta_j)\exp\left\{\sum_{i=1}^{k}\tau_i h_i(y_j;\theta_j)\right\} f(y_j;\theta_j).$$

This structure allows for 'smooth' departures from the assumed distribution controlled by the vector parameter, $\boldsymbol{\tau} = [\tau_1,\ldots,\tau_k]^T$ acting on the elements of the set, $\{h_i(y;\theta)\}$, of polynomials up to order k which are orthonormal on the assumed distribution. The normalizing constant, $C(\boldsymbol{\tau},\theta_j)$, simply ensures that $f_k(y_j;\boldsymbol{\tau},\theta_j)$ is correctly scaled to provide a valid probability density function.

When $\boldsymbol{\tau} = \mathbf{0}$, this smooth alternative collapses to the original response variable distribution. Thus a test of $H_0 : \boldsymbol{\tau} = \mathbf{0}$ against $H_A : \boldsymbol{\tau} \neq \mathbf{0}$ can reasonably be considered a test of the distributional assumption in a generalized linear model.

In [3], a score test statistic has been derived that can be expressed as a sum of squares of several contributing components:

$$\hat{S}_k = \frac{\hat{V}_1^2}{\hat{\omega}^2} + \hat{V}_2^2 + \ldots + \hat{V}_k^2$$

where

$$\hat{V}_i = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} h_i(y_j;\hat{\theta}_j).$$

The ith component involves the sum over the data of the ith order polynomial from the orthonormal sequence used in the construction of the smooth alternative distribution. The first component also contains a

term

$$\omega^2 = 1 - \frac{\mathbf{1}^T \boldsymbol{H} \mathbf{1}}{n}$$

which is related to the hat matrix, $\boldsymbol{H}$, obtained from the model estimation process.

Large values of $\hat{S}_k$ provide evidence against H_0. Asymptotically, the components $\hat{V}_1^2/\hat{\omega}^2$, $\hat{V}_2^2$, etc can each be expected to follow the $\chi^2_{(1)}$ distribution and $\hat{S}_k$ the $\chi^2_{(k)}$ distribution. In practice this has not proved a good enough approximation for common sample sizes and so a parametric bootstrap process is recommended to estimate p-values.

3. Applying the Smooth Test

In deriving this test of the distributional assumption, the linear predictor and the link function are assumed to be correctly specified. If this is not true then a large value of the test statistic may be caused by a mismatch between the data and these other components of the generalized linear model rather than an inappropriate response distribution. Similar issues arise with other tests that are used to assess generalized linear models. For example, the well-known deviance statistic is derived as a likelihood ratio test statistic comparing the fitted model with a saturated model having a linear predictor with as many parameters as there are covariate patterns. This provides the best possible fit to the observed data – assuming that the specified response distribution and link function are correct. If this is not true, then a large value of the deviance statistic may indicate a problem with the assumed distribution or link function rather than the linear predictor. Similarly, a model that 'fails' a goodness-of-link test may really have a problem with the assumed distribution or linear predictor and not the link function.

Can we ever truly diagnose the problem with a poorly fitting model? Clearly all such tests need to be carefully interpreted. There are many different ways that a model can be misspecified, some of which are very difficult to distinguish from each other. The smooth testing approach is not a panacea. In addition to providing a reliable test of the distributional assumption however, the individual components can be considered as test statistics in their own right. This can provide useful diagnostic information about the nature of any lack of fit detected.

4. Power Study

4.1. Logistic Regression

Figure 1 shows the results of a simulation study for logistic regression with a **misspecified linear predictor**. In this example, the fitted model was

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1$$

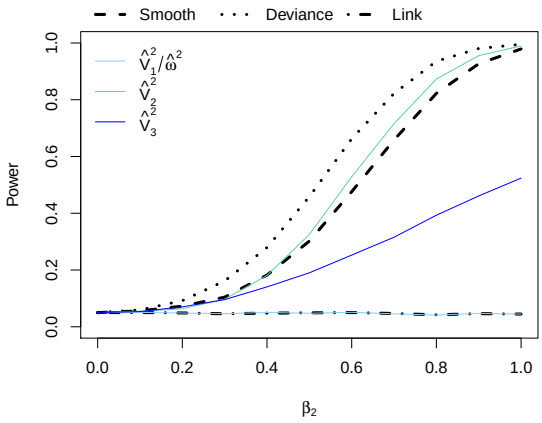

Figure 1: Power to detect a misspecified linear predictor in simulated logistic regression data.

but the true model used to simulate the data was

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

A fixed covariate pattern was used for each simulation with 25 groups corresponding to x_1 taking values $-1, -0.5, 0, 0.5, 1$ and x_2 taking values $-1.2, -0.7, -0.2, 0.3, 0.8$. There were $m = 30$ trials in each group. These two models coincide when $\beta_2 = 0$. The misspecification increases as β_2 increases (horizontal axis).

100000 simulations were conducted for $\beta_2 = 0$ to characterize the null distribution of each test statistic and 20000 simulations for each of the other β_2 values to characterize the alternative distributions. The $\alpha = 5\%$ critical value from the null distribution was used to define the rejection region and thus determine the probability of the null hypothesis being rejected (power to detect the misspecification) which is plotted on the vertical axis.

Three test statistics have been considered here: the deviance statistic, the smooth test statistic of order 3 and a link test statistic (see Appendix A). For all statistics used, the powers were based on simulated distributions and not on approximate sampling distributions. In this first example, the deviance performs best in detecting this particular kind of misspecification of the linear predictor. But the smooth test still performs reasonably well and the link test is essentially useless here. The performance of the $\hat{S}_k$ statistic is a compromise between the performance of the individual components which can also be considered separately. In this case: the first component is almost exactly matching the performance of the goodness of link test; the second component has good power and drives the performance of the overall test statistic and the third component is not particularly useful. The components correspond roughly to moments and so the second component is suggesting that the variance in the data is not

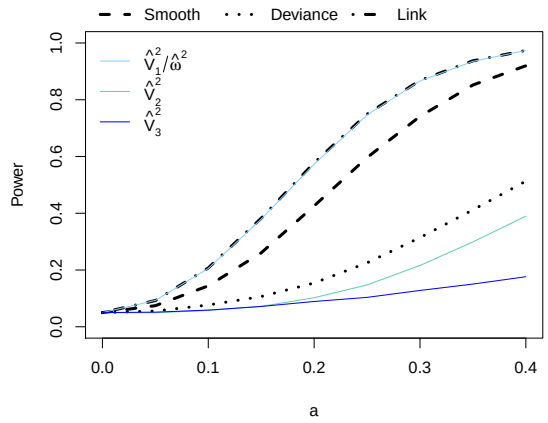

Figure 2: Power to detect a misspecified link function in simulated logistic regression data.

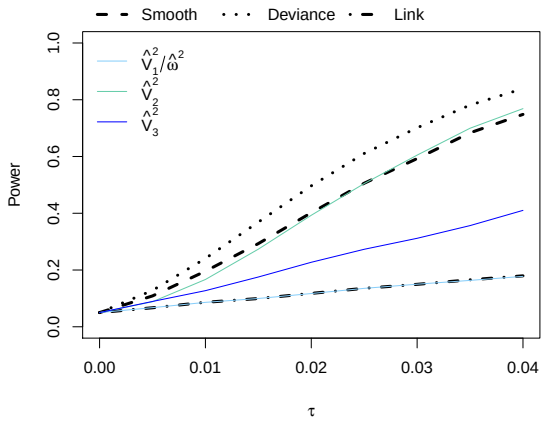

Figure 3: Power to detect a misspecified response distribution in simulated logistic regression data.

well modelled. This makes sense. A covariate is missing and so the stochastic part of the model is trying to cope with additional variation that should really have been explained by the linear predictor.

Figure 2 shows the results for a **misspecified link function** where the fitted model was

$$\pi(\eta) = \frac{e^{\eta}}{1+e^{\eta}} \qquad \log\left(\frac{\pi}{1-\pi}\right) = \eta = \beta_0 + \beta_1 x_1$$

but the data was simulated using a generalization of the logit link function (see Appendix B):

$$\pi(\eta) = \frac{e^{h(\eta;a)}}{1+e^{h(\eta;a)}}. \qquad (1)$$

The parameter a plotted along the horizontal axis controls the amount of misspecification with zero again representing no misspecification. Other simulation details are the same as in the first example.

Unsurprisingly, it is the goodness of link test that performs best here as this is the kind of problem it is designed to detect. However, the smooth test still performs well. Looking at the individual components, the first component is again matching the performance of the goodness of link test and is driving the performance of the overall test statistic in detecting this kind of misspecified model. The first component is correctly indicating that the problem is in how the mean of the data is being modelled. The second and third components aren't useful in this case.

Figure 3 shows the results for a **misspecified response distribution** where a binomial distribution is specified when fitting the model but the data was simulated using a beta-binomial distribution where the responses Y_j are $\mathrm{B}(m_j, \pi_j^*)$ for π_j^* independently distributed as beta random variables on $(0,1)$ with $E[\pi_j^*] = \pi_j$ and $\mathrm{Var}(\pi_j^*) = \tau\pi_j(1-\pi_j)$.

Again the parameter plotted along the horizontal axis, τ in this case, controls the amount of misspecification with zero representing no misspecification. The

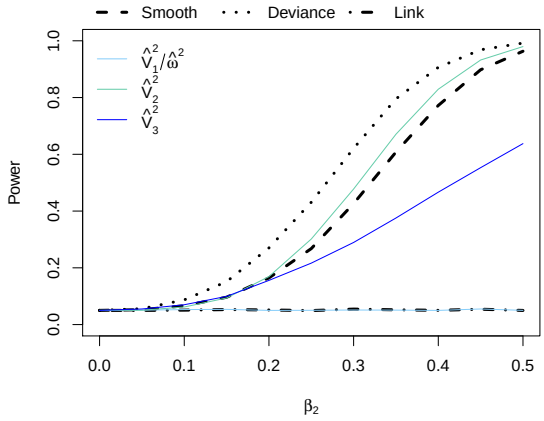

Figure 4: Power to detect a misspecified linear predictor distribution in simulated Poisson regression data.

deviance test performs best in detecting this particular type of misspecification, with the smooth test again performing reasonably well and the goodness of link test poorly. The story with the components is again similar with the first component matching the performance of the goodness of link test and the second component indicating correctly that the variance is not being modelled correctly in this example.

4.2. Poisson Regression

In Figure 4, the simulation scenario is the same as for Figure 1 except that the linear predictor is set to $\log\mu$ where $Y_j \sim \mathrm{P}(\mu_j)$. The performance of the smooth test statistic and components in detecting this type of misspecified linear predictor in Poisson regression can be seen to be very similar to that already discussed for logistic regression.

In Figure 5, a Poisson distribution is specified when fitting the model but the data was simulated using a negative binomial distribution with $\log\mu_j = \eta_j$ and variance $\mu_j + \tau\mu_j^2$. As in the similar logistic regression example, the deviance is more powerful in detecting the misspecification but the smooth test performs rea-

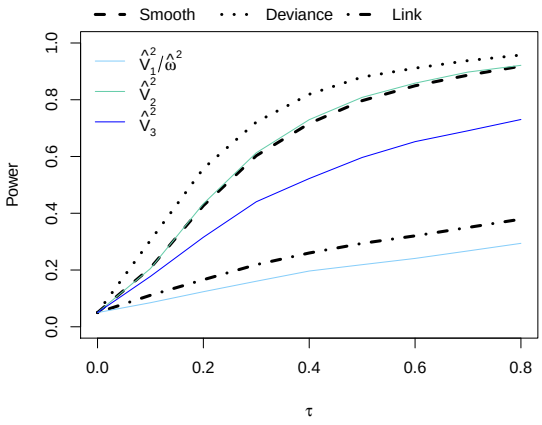

Figure 5: Power to detect a misspecified response distribution in simulated Poisson regression data.

sonably and the second component correctly indicates that the problem is in how the variance of the data is being modelled.

5. Conclusions

A smooth test for assessing the distributional assumption in generalized linear models has been derived in [3] and applied here to Poisson and logistic regression models fitted to simulated data. While not always the most powerful test, it appears to perform quite well in detecting lack of fit even when the misspecification is in the link function or the linear predictor rather than the response distribution. Interpretation of the components provides additional diagnostic information.

A. Goodness of Link Test

There are a number of tests described in the literature for testing the adequacy of the link function in a generalized linear model. Many of these are specific to a particular link function. The goodness of link test used in this paper is more generally applicable and is equivalent to the `linktest` function provided in STATA [4].

The $\hat{\eta} = \boldsymbol{X}\hat{\beta}$ term from the fitted model and a $\hat{\eta}^2$ term are used as the predictors of the original response variables in a new model. The $\hat{\eta}$ term contains all the explanatory information of the original model. If there is a misspecified link the relationship between $\hat{\eta}$ and $g(\overline{y})$ will be non-linear and the $\hat{\eta}^2$ term is likely to be significant. The difference in deviance between these two models has been used as the link test statistic in this study.

B. Generalized Logit Function

Expressed as an inverse link function, a generalization of the logit function is described by [5] in the same form as Eq. (1) but using a function $h(\eta; \alpha_1, \alpha_2)$ where the two shape parameters, α_1 and α_2, separately control the left and right tails. $\alpha_1 = \alpha_2$ gives a symmetric probability curve $\pi(\eta)$ with the logistic model as the special case $\alpha_1 = \alpha_2 = 0$. The function $h(\eta; a)$ used in Eq. 1 corresponds to $a = -\alpha_1 = \alpha_2$. This gives an asymmetric probability curve that according to [5] corresponds to a Box-Cox power transform.

References

[1] J. Neyman. Smooth tests for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20:149–199, 1937.

[2] J. C. W. Rayner, O. Thas, and D. J. Best. *Smooth tests of goodness of fit: Using R*. Oxford University Press, 2nd edition, 2009.

[3] Paul Rippon. Application of smooth tests of goodness of fit to generalized linear models. *Unpublished PhD thesis.*, 2011.

[4] StataCorp. *Stata Base Reference Manual, Release 9*. Stata press, 2005.

[5] Therese A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431, 1988.