2011

Assessing Poisson and Logistic Regression Models Using Smooth Tests

Paul Rippon
University of Newcastle

John Rayner
University of Newcastle

Publication Details

Rippon, Paul; and Rayner, John, Assessing Poisson and Logistic Regression Models Using Smooth Tests, Proceedings of the Fourth Annual ASEARC Conference, 17-18 February 2011, University of Western Sydney, Paramatta, Australia.

Assessing Poisson and Logistic Regression Models Using Smooth Tests

Abstract

The smooth testing approach described in [2] has been used to develop a test of the distributional assumption for generalized linear models. Application of the test to help assess Poisson and logistic regression models is discussed. Power is compared to other common tests.

Keywords

generalized linear models, goodness of fit, logistic regression, Poisson regression

Publication Details

Rippon, Paul; and Rayner, John, Assessing Poisson and Logistic Regression Models Using Smooth Tests, Proceedings of the Fourth Annual ASEARC Conference, 17-18 February 2011, University of Western Sydney, Paramatta, Australia.

Assessing Poisson and Logistic Regression Models Using Smooth Tests

Paul Rippon, J.C.W. Rayner

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA

Abstract

The smooth testing approach described in [2] has been used to develop a test of the distributional assumption for generalized linear models. Application of the test to help assess Poisson and logistic regression models is discussed. Power is compared to other 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 the 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 $x_{j1} = 1$ in the first column of $\boldsymbol{X}$.

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 $\boldsymbol{\tau} = \boldsymbol{0}$, this smooth alternative collapses to the original response variable distribution. Thus a test of $H_0 : \boldsymbol{\tau} = \boldsymbol{0}$ against $H_A : \boldsymbol{\tau} \neq \boldsymbol{0}$ 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.

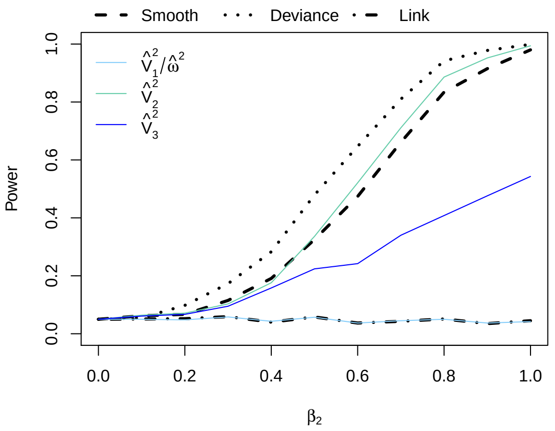

Figure 1: Power to detect a misspecified linear predictor in simulated logistic regression data.

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

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

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).

5000 simulations were conducted for $\beta_2 = 0$ to characterize the null distribution of each test statistic and 1000 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

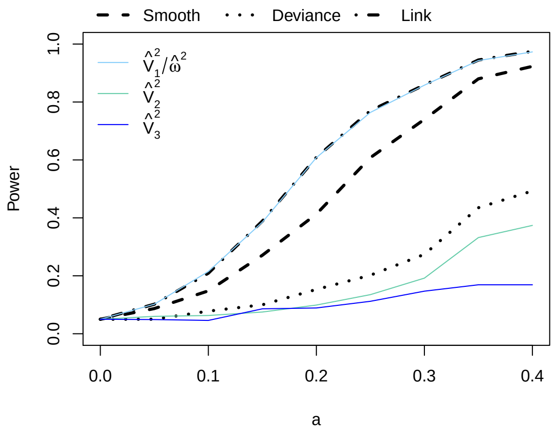

Figure 2: Power to detect a misspecified link function in simulated logistic regression data.

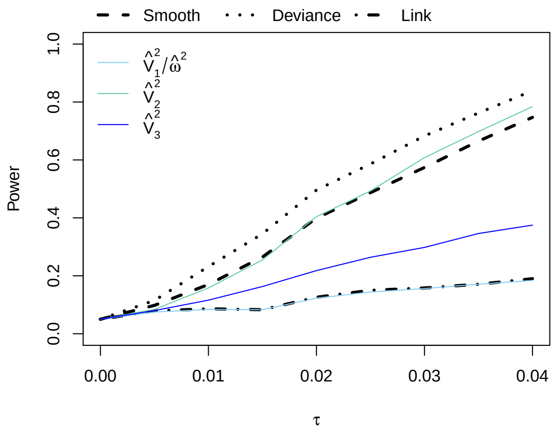

Figure 3: Power to detect a misspecified response distribution in simulated logistic regression data.

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 indicating that the variance in the data is not 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 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

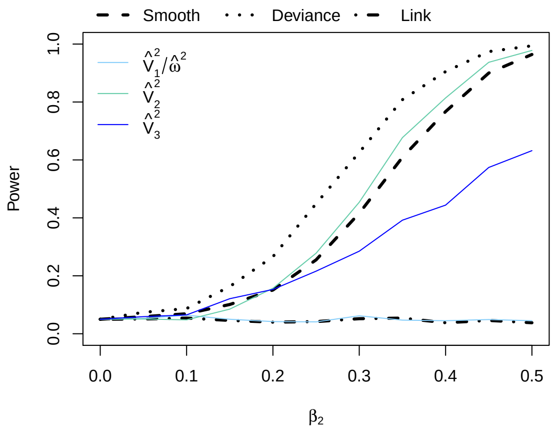

Figure 4: Power to detect a misspecified linear predictor distribution in simulated Poisson regression data.

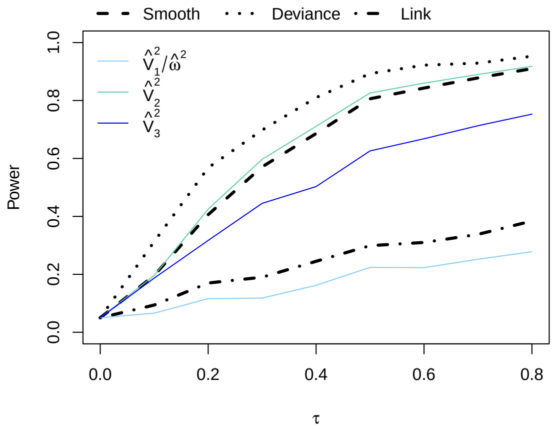

Figure 5: Power to detect a misspecified response distribution in simulated Poisson regression data.

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 reasonably 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 and applied 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.