NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

05-17

A Tutorial on Smooth Tests of Fit for Poisson and Logistic Regression

D.J. Best and J.C.W. Rayner

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

A Tutorial on Smooth Tests of Fit for Poisson and Logistic Regression

D.J. Best & J.C.W. Rayner

1. Introduction

Smooth tests have been used to assess the goodness of fit of many statistical distributions. See, for example, the illustrations given in Rayner et al. (2009). More recently they have been used to assess the response distribution assumption in Poisson and logistic regression. See Rippon and Rayner (2011). In the former situation where the data are assumed to be identically distributed, the smooth test for n data points $y_1, \ldots, y_n$ is based on the statistic

$$S_k = \sum_{i=1}^{k} V_i^2$$

in which $V_i = \sum_{j=1}^{n} h_i(y_j;\mu)/\sqrt{n}$, $\{h_i(y;\ \mu\}$ is the set of polynomials orthonormal on the distribution tested for that depends on the parameter μ and in which k is the order of the test, predetermined by the data analyst. The V_i^2 are said to be components of the smooth test statistic S_k. Sometimes these components are also the basis of smooth tests in their own right. In the following we will look at the component V_2^2 which Ozonur et al. (2016) shows performs well as a test of fit for the response variable in Poisson regression.

In the regression situation the responses are not identically distributed due to the regression predictor variables $X_1, \ldots, X_m$ say, and now $V_2 = \sum_{j=1}^{n} h_2(y_j;\mu_j)/\sqrt{n}$. Note the extra j subscript on μ. In the Poisson regression the μ_j are both

- expected values of the Y_j and
- the Poisson parameter of the distribution of the Y_j which may vary for each j.

For the Poisson regression case, if $t_j = y_j - \mu_j$ and for the jth observation

$$h_2(y_j;\mu_j) = (t_j^2 - t_j - \mu_j)/\sqrt{(2\mu_j^2)},\ \mu_j \neq 0.$$

For the logistic regression case we suppose there are n different sets of predictors or covariates, with the jth having N_j observations and probability of success p_j. Then for $j = 1, \ldots, n$

$$h_2(y_j;\mu_j) = \frac{(y_j - N_j p_j)^2 + (2p_j - 1)y_j - N_j p_j^2}{p_j(1-p_j)\sqrt{2N_j(N_j-1)}}.$$

Observe that if $N_j = 1$ or if $p_j = 0$ or 1 then we can use, say, $N_j = N_j + \delta$, $p_j = p_j + \delta$ for δ small, say $\delta = 0.0001$. Similarly for Poisson regression if $\mu_j = 0$ take $\mu_j = \delta$.

2. Poisson regression example

The following example is from Draper and Smith (1998, p.406). Quoting these authors

> "Table 1 shows a set of data on reported occurrences of a communicable disease in two areas of the country at ten 2 month intervals, 2, 4, …, 20. There are 20 data points, so that $i = 1, 2, \ldots, 20$ below.
>
> We assume that the occurrences Y_j are Poisson variables with $E[Y_j] = \mu_j$ and that
>
> $$\ln \mu_j = \beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2}$$
>
> is the model under consideration. We take $X_{j1} = \ln(2j)$ corresponding to ln(month indicator) and $X_{j2} = 0$ for observations from area A and $X_{j2} = 1$ for observations in area B. Other choices of X_{j1} and X_{j2} are possible."

Table 1. Communicable disease data

Y_j	X_{j1}	X_{j2}	$\hat{\mu}_j$
8	0.693	0	6.999
8	1.386	0	9.098
10	1.792	0	10.609
11	2.079	0	11.826
14	2.303	0	12.873
17	2.485	0	13.790
13	2.639	0	14.418
15	2.773	0	15.379
17	2.890	0	16.075
15	2.996	0	16.733
14	0.693	1	13.178
19	1.386	1	17.130
16	1.792	1	19.975
21	2.079	1	22.267
23	2.303	1	24.237
27	2.485	1	25.965
28	2.639	1	27.523
29	2.773	1	28.955
33	2.890	1	30.266
31	2.996	1	31.505

Using Poisson regression software

$$\ln \hat{\mu}_j = 1.684 + 0.3784\, X_{j1} + 0.6328\, X_{j2}$$

and the usual deviance statistic is 3.126 which is not significant using the χ^2_{17} approximation. Thus the y_j and μ_j values would seem to be fairly close. However V_2^2 = 7.150 and using the χ^2_1 approximation the value of V_2^2 is highly significant. Thus the responses y_j do *not* appear to be Poisson distributed and so fitting a Poisson regression would not be sensible. For this example note that n = 20.

We also note that V_2^2 is more sensitive to dispersion alternatives than the usual deviance statistic and this is the probable cause of the differing conclusions.

3. Logistic regression example

Dobson (2002, p.119) fits a logistic regression to the data in Table 2.

Table 2. Beetle mortality data

Dose x_j ($\ln_{10}CS_2$ mgl $^{-1}$)	Number of beetles N_j	Number killed y_j	$N_j \hat{p}_j$
1.6907	59	6	3.458
1.7242	60	13	9.842
1.7552	62	18	22.451
1.7842	56	28	33.898
1.8113	63	52	50.096
1.8369	59	53	53.291
1.8610	62	61	59.222
1.8839	60	60	58.743

Using logistic regression software

$$\ln\left(\frac{\hat{p}_j}{1-\hat{p}_j}\right) = -\,60.72 + 34.72\, X_{j1}$$

and the usual deviance statistic takes the value 11.23 which is not significant at the 0.05 level based on the χ^2_6 approximation. Thus the y_j and $N_j p_j$ values seem reasonably close. Also V_2^2 = 0.472 and using the χ^2_1 approximation this value of V_2^2 is also not significant. Thus the responses y_j appear to be binomially distributed and so fitting a logistic regression is sensible. For this example n is the number of different covariate combinations and is 8.

4. Parametric bootstrap p-values and powers

We discuss this for the Poisson regression example above but the approach is the same for logistic regression. Rippon and Rayner (2011) suggest the χ_1^2 approximation may not be good and that p-values should be found using a parametric bootstrap.

To do so, for each observation $y_1, \ldots, y_n$ we obtain a random value from a Poisson $\hat{\mu}_j$ distribution. For this new data set a new value of V_2^2 is calculated. This new V_2^2 is compared with V_2^2 for the original data. If it is greater than or equal to the original value, 7.150, a *counter* initialised at zero is increased by 1. This process is repeated a large number of times, *nsim* say. The ratio *counter*/*nsim* is a parametric bootstrap p-value estimate. Using the χ_1^2 approximation the p-value for the V_2^2 test statistic is 0.007; using *nsim* = 10,000 a bootstrap p-value estimate is 0.004. The difference in p-values for this example is not important.

In Table 3 below powers are based on 10,000 simulations of the y_j values based on the alternative distribution. P-values for each of the statistics for the regressions based on these 10,000 simulations are also based on a further 10,000 simulations, assuming the y_j now have a Poisson distribution. A p-value less than 0.05 is counted as significant and the fraction of these p-values which are significant is the estimated power.

Table 3. Power Comparison

Model	Alternative	T^2	V_2^2
1	NB(τ = 0.4)	0.20	0.26
1	PM(δ = 0.15)	0.78	0.73
2	NB(τ = 0.4)	0.18	0.23
2	PM(δ = 0.3)	0.62	0.53

In Table 3 we looked at negative binomial and Poisson mixture alternatives to the Poisson response distribution. Such overdispersed alternatives are thought to be the most likely to occur in practice. Following Spinelli et al. (2002) two models for n data points (from the large number of possibilities) were considered. In model 1, $\mu_j = \exp(2.6 + 2x_j)$ with x_j = 5*0.0, 5*0.5 and 5*1.0 and in model 2, $\mu_j = \exp(3x_j)$ with the same x_j. Here n = 15. The powers in Table 3 are for a significance level $\alpha = 0.05$. The negative binomial alternative had mean μ_j and variance $\mu_j(1 + \tau)$ with $\tau = 0.4$ for both models 1 and 2. The Poisson mixture was composed of two equiprobable Poisson distributions with means $\mu_j - \delta\mu_j$ and $\mu_j + \delta\mu_j$ for both models 1 and 2. For model 1, δ is 0.15 and for model 2, δ is 0.3. The square of the Dean

(1992) statistic $T^2 = \{\sum_{j=1}^{n}[(y_j - \hat{\mu}_j)^2 - y_j]\}^2 / (2\sum_{j=1}^{n}\hat{\mu}_j^2)$ is compared to V_2^2 in Table 3. Large values of T^2 and V_2^2 are considered significant. T is P_B and V_2 is P_C in Dean (1992).

Table 3 shows that neither T^2 nor V_2^2 is always the more powerful statistic although T^2 is often considered a good test for overdispersion. The relative advantages of powers in Table 3 can be repeated for other τ, δ, n and α.

References

Dean, C. (1992). Testing for overdispersion in Poisson and logistic regression models. *Journal of the American Statistical Association*, **87**, 451-457.

Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*, 2nd ed., Boca Raton: Chapman & Hall.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd ed., New York: Wiley.

Ozonur, D., Akdur, H. and Bayrak, H. (2016). Comparisons of tests of distributional assumption in Poisson regression model. *Communications in Statistics - Simulation and Computation*, DOI: 10.1080/03610918.2016.1202267.

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

Rippon, P. (2013). Application of Smooth Tests of Goodness of Fit to Generalized Linear Models. PhD Thesis, University of Newcastle, Australia.

Rippon, P. and Rayner, J.C.W. (2011). Assessing Poisson and logistic regression models using smooth tests. *Proceedings of the Fourth Annual Applied Statistics Education and Research Collaboration (ASEARC) Research Conference, February 17—18, 2011*: Parramatta, Australia.

Spinelli, J.J., Lockhart, R.A. and Stephens, M.A. (2002). Tests for the response distribution in a Poisson regression model. *Journal of Statistical Planning and Inference*, **108**, 137-154.