ECONOMICS BULLETIN

"A regression error specification test (RESET) for generalized linear models".

Sunil Sapra
Department of Economics Statistics, California State University, Los Angeles

Abstract

Generalized linear models (GLMs) are generalizations of linear regression models, which allow fitting regression models to response data that follow a general exponential family. GLMs are used widely in social sciences for fitting regression models to count data, qualitative response data and duration data. While a variety of specification tests have been developed for the linear regression model and are routinely applied for testing for misspecification of functional form, omitted variables, and the normality assumption, such tests and their applications to GLMs are uncommon. This paper develops a regression error specification test (RESET) for GLMs as an extension of the popular RESET for the linear regression model (Ramsey (1969)). Applications of the RESET to three economic data sets are presented and the finite sample power properties are studied via a Monte Carlo experiment.

Citation: Sapra, Sunil, (2005) ""A regression error specification test (RESET) for generalized linear models"..." *Economics Bulletin*, Vol. 3, No. 1 pp. 1−6
Submitted: November 5, 2004. **Accepted:** January 10, 2005.
URL: http://www.economicsbulletin.com/2005/volume3/EB−04C50033A.pdf

I. Introduction

Inference procedures for regression models assume that the response variable y follows the normal distribution. There are, however, many situations in social sciences where this assumption fails to hold. Common examples are count data models, qualitative response models, and duration data models. The data may involve such variables as the number of trips to a doctor's office during a year, choice of a mode of transportation, decision to purchase an item at a given time, or the duration of unemployment or a strike. The generalized linear models (GLMs) deal with such situations involving non-normal data.

Several specification tests have been developed for testing various types of misspecification in linear and nonlinear regression models including tests for normality of the error distribution, omitted variables, and misspecification of the functional form. Use of such tests for GLMs, however, is not too common with few exceptions such as a test for overdispersion in count data models and specification tests for qualitative response models developed by Davidson and MacKinnon (1984). This paper extends the popular regression specification error test (RESET) (Ramsey (1969)) to GLMs. The RESET test was developed to detect omitted variables and incorrect functional form in the linear regression model. The paper applies the RESET test to three different economic data sets and studies the power properties of the test via a Monte Carlo experiment.

This paper is organized as follows. Section II presents the GLM and develops two versions of the RESET test for GLMs. Section III presents three applications of the test to count and qualitative response data. Section IV presents the results of a Monte Carlo experiment on the power properties of the test. Section V provides some concluding remarks.

II. The GLM and the RESET test for the GLM

The GLM is a generalization of linear regression model to fit data in situations where the response variable follows a distribution, which is a member of the linear exponential family.

The probability density function for the GLM is

$$f(y;\theta,\phi)=\exp[\{y\theta-b(\theta)\}/a(\phi)+c(y,\phi)], \qquad (1)$$

where $a(.), b(.),$ and $c(.)$ are known functions. The parameter θ is a natural location parameter and ϕ is called a dispersion parameter.

The GLM has the following components (see Gill (2001), McCullagh and Nelder (1989) and Myers et al. (2002)).

1. *The Stochastic Component* : $y_1, y_2,..., y_n$ are indpendent response observations with means $\mu_1, \mu_2,..., \mu_n$ respectively.

2. *The Systematic Component* : The model is constructed from the linear predictor

$$\eta_{\mathrm{i}}=\beta' x_i=\beta_0+\sum_{i=1}^{k}\beta_i x_i. \qquad (2)$$

3. *The Link Function* : The link function η_{i} links the stochastic and the systematic parts and is defined as

$$\eta_{\mathrm{i}}=g(\mu_i), i=1,2,...,n, \qquad (3)$$

where $g(.)$ is a monotonic differentiable function. Therefore,

$$\mu_{\mathrm{i}}=g^{-1}(\eta_i)=g^{-1}(\beta' x_i), i=1,2,...,n. \qquad (4)$$

4. The variance σ_{i}^2 of y_i $(i=1,2,...,n)$ is a function of the mean μ_{i}.

RESET Tests for GLMs

Our proposed RESET tests compare a GLM with no higher order terms with a GLM with higher order terms. Specifically, RESET1 compares a GLM with no higher order terms with a GLM with the second power of the predicted link function and RESET2 compares the former with a GLM with the third power of the predicted link function. Consider the following three GLMs with the same stochastic component but different systematic components.

$GLM1: \eta_{1i} = \beta' x_i.$

$GLM2: \eta_{2i} = \beta' x_i + \gamma_1 \hat{\eta}_{1i}^2.$

$GLM3: \eta_{3i} = \beta' x_i + \gamma_1 \hat{\eta}_{1i}^2 + \gamma_2 \hat{\eta}_{1i}^3$

where $\hat{\eta}_{1i} = \hat{\beta}' x_i$ and $\hat{\beta}$ is the maximum likelihood estimator (MLE) of β under GLM 1.

We consider the following RESET tests.

$RESET1$: Compare $GLM1$ and $GLM2$ and test $H_0 : \gamma_1 = 0$ against $H_1 : \gamma_1 \neq 0$.

$RESET2$: Compare $GLM1$ and $GLM3$ and test $H_0 : \gamma_1 = \gamma_2 = 0$ against H_1 : At least one of the coefficients γ_1 and γ_2 is not equal to 0.

Let L_1, L_2 and L_3 denote the log - likelihood functions evaluated at the MLEs under $GLM1$, $GLM2$, and $GLM3$ respectively. Then the test statistics for $RESET1$ and $RESET2$ are respectively

$\lambda_1 = -2(L_1 - L_2) \sim \chi^2(1)$ under $H_0 : \gamma_1 = 0$ and

$\lambda_2 = -2(L_1 - L_3) \sim \chi^2(2)$ under $H_0 : \gamma_1 = \gamma_2 = 0.$

III. Empirical Examples

In this section, we present three empirical applications of *RESET*1 and *RESET*2 tests.

III a. Application to Capital Punishment Data

The data on capital punishment in 17 states in the US are from Gill (2001). The outcome variable is the number of times capital punishment is implemented in on a state level in the United States for the year 1997. The explanatory variables are median per capita income in dollars, the percent of the population classified as living in poverty, the percent of Black citizens in the population, the rate of violent crimes per 100,000 residents for the year before (1996), a dummy variable to indicate whether the state is in the South, and the proportion of the population with a college degree of some kind. The Poisson regression model with the log link was fitted to the data. The results are summarized in Table 1.

Table 1. *Capital Punishment in the United States*

Variables	Coefficient	Standard Error
Intercept	-6.8014798	4.1468731
Median Income	.0002611	.000519
Percent Poverty	.077818	.0794026
Percent Black	-.0949311	.0229193
Log(Violent Crime)	.2969349	.43751757
South	2.3011833	.4283838
Degree Proportion	-18.722068	4.2839793

*RESET*1*:* λ_1 = 4.62914, *RESET*2*:* λ_2 = 7.36024

The values of chi-squared statistics for both *RESET*1 and *RESET*2 indicate that the squares and cubes of the predicted values of the link function are significant at 5% and 10% significance levels. The null hypothesis is rejected under both tests indicating that the Poisson model with log link is inadequate.

III b. Application to Transportation Data

The data on automobile and public transportation travel times and the alternative chosen for 21 individuals are from Ben-Akiva and Lerman (1985). The dependent variable y = 1 if automobile transportation is chosen and 0 if public transportation is chosen. The explanatory variable is x = bus time - auto time. A logistic regression model was fitted to the data. The results are presented in Table 2.

Table 2. *Transportation Data*

Variables	Coefficient	Standard Error
Intercept	-.23757544	.75047663
X	.055310983	.020642279

*RESET*1*:* $\lambda_1 = 1.020384$, *RESET*2*:* $\lambda_2 = 12.183274$

The value of chi-squared statistics for *RESET*1 is too small and the predicted value of the link function is insignificant at 5% and 10% significance levels indicating that the logit model is adequate. However, *RESET*2 indicates that the squares and cubes of the predicted values of the link function are significant at 5% and 10% significance levels suggesting that the logit model is inadequate.

III c. Application to Multiple Bids Data

The data on multiple bids are from Jaggia and Thosar (1993) and are also analyzed in Cameron and Trivedi (1998). The data consist of 126 observations on U.S. firms that were targets of tender offers during the period from 1978 through 1985 and were taken over within 52 weeks of the initial offer. As in these studies, the dependent variable is the number of bids after the initial bid (NUMBIDS) received by the target firm and the explanatory variables are LEGLREST, REALREST, FINREST, WHITEKNT, BIDPREM, INSTHOLD, SIZE, SIZESQ, and REGULATN. A Poisson regression model with logarithmic link was fitted to the data. The results of Poisson MLE are presented in Table 3.

Table 3. *Multiple Bids Data**

Variables	Coefficient	Standard Error
Intercept	.9860599	.53392014
LEGLREST	-.6776959	.37673724
REALREST	-.36199125	.42432924
FINREST	.17850260	.060022105
WHITEKNT	.26014637	.15095939
BIDPREM	-.19565974	.19263088
INSTHOLD	.074030059	.21652194
SIZE	-.029439199	.16056816
SIZESQ	.4813821684	.15886982
REGULATN	-.00756935	.00312170

RESET 1: $\lambda_1 = 1.1154$, *RESET 2:* $\lambda_2 = 1.191$

* The coefficient estimates and standard errors reported are based on Poisson MLE, while those reported in Cameron and Trivedi (1998) (p. 148) are based on Poisson pseudo MLE.

The values of the chi-squared statistics are too small under both *RESET*1 and *RESET*2 and the predicted values of the link function are insignificant at 5% and 10% significance levels. The null hypothesis is not rejected under either test indicating that the Poisson model with log link is adequate.

IV. Monte Carlo Experiment

100 samples of sizes 50, 100 and 200 on the variable y were generated according to the Poisson law with log link based on the mean functions

$$\mu_i = \exp(2 + 3x_i + 5x_i^2), \qquad (5)$$

$$\mu_i = \exp(2 + 3x_i + 5x_i^2 + 6x_i^3) \qquad (6)$$

for *RESET*1 and *RESET*2 respectively.

The sample on the right-hand side variable *x* was generated according to the uniform law *U*(0,2) and held fixed once it was generated. The model under the null hypothesis was the incorrect Poisson model with the link function

$$\eta_{1i} = \ln(\mu_{1i}) = \beta_1 + \beta_2 x_i, \qquad (7)$$

which does not include second and third degree terms. The models under the alternative hypotheses under tests *RESET* 1 and *RESET* 2 were Poisson with the link functions

$$\eta_{2i} = \ln(\mu_{2i}) = \beta_1 + \beta_2 x_i + \gamma_1 \hat{\eta}_{1i}{}^2, \qquad (8)$$

$$\eta_{3i} = \ln(\mu_{3i}) = \beta_1 + \beta_2 x_i + \gamma_1 \hat{\eta}_{1i}{}^2 + \gamma_2 \hat{\eta}_{1i}{}^3 \qquad (9)$$

respectively.

The results on the power properties are presented in table 4.

Table 4. *Power properties of RESET*1 *and RESET*2 *Tests for* α *=.05*

Sample size (n)	Estimated power of *RESET*1	Estimated Power of *RESET* 2
50	.76	.61
100	.97	.86
200	.99	.99

Table 5. *Power properties of RESET*1 *and RESET*2 *Tests for* α *=.01*

Sample size (n)	Estimated power of *RESET*1	Estimated Power of *RESET* 2
50	.53	.44
100	.84	.78
200	.97	.97

Table 6. *Power properties of RESET*1 *and RESET*2 *Tests for* α *=.10*

Sample size (n)	Estimated power of *RESET*1	Estimated Power of *RESET*2
50	.88	.69
100	.99	.95
200	.99	.99

At all of the significance levels in the tables above, the powers of both *RESET*1 and *RESET*2 tests tend to increase as the sample size increases from 50 to 200. Furthermore, the power of each test increases as the significance level increases. It is also clear that for each significance level and sample size, *RESET*1 has higher power than *RESET*2. This finding simply reflects the well-known result that the power of a chi-squared test at any given significance level is a strictly decreasing function of degrees of freedom (Das Gupta and Perlman (1974)). Finally, the power of *RESET*2 approaches that of *RESET*1 as the sample size increases.

V. CONCLUSION

Motivated by specification tests for testing for functional form and omitted variables in linear regression model, this paper has developed two versions of the regression error specification tests (RESET) for GLMs. The tests were applied to some data sets from economics and other social sciences. Our limited simulation results suggest that the RESET tests for GLMs have reasonable power properties in medium to large samples. These tests are computationally convenient and require only the predicted value of the link function and maximum values of the log-likelihood functions under the null and alternative hypotheses, which can be easily computed using common econometric and statistical software packages. Applications of such tests to count, qualitative response and duration data models in the GLM family should become routine given their computational convenience and good power properties.

REFERENCES

Ben-Akiva, M. and Lerman, S. (1985) *Discrete Choice Analysis*, MIT Press, Cambridge, MA.

Cameron, A. C. and Trivedi, P. K. (1998) *Regression Analysis of Count Data*, Cambridge University Press, New York.

Das Gupta, S. and Perlman, M. D. (1974) "Power of the noncentral F-test: Effect of additional variates on Hotelling's T^2 test", *Journal of the American Statistical Association*, **69**, 174-80.

Davidson, R. and J. G. MacKinnon (1984) "Convenient specification tests for logit and probit models", *Journal of Econometrics*, **25**, 241-62.

Gill, J. (2001) *Generalized Linear Models, A Unified Approach*, Sage University Press, California.

Jaggia, S. and S. Thosar (1993) “Multiple Bids as a Consequence of Target Management Resistance: A Count Data Approach”, *Review of Quantitative Finance and Accounting*, **3**, 447-457.

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

Myers, R. H., Montgomery, D. C. and Vining G. G. (2002) *Generalized Linear Models with Applications in Engineering and the Sciences,* Wiley, New York.

Ramsey, J. B. (1969) “Tests for specification errors in classical least-squares regression analysis”, *Journal of the Royal Statistical Society, Series B*, **31**, 350-71.