The Canadian Journal of Statistics
Vol. 28, No. 2, 2000, Pages 439–445
La revue canadienne de statistique

Effects of omitting a covariate in Poisson models when the data are balanced

Martin R. PETERSEN and James A. DEDDENS

Key words and phrases: Generalized linear model; maximum-likelihood estimation; Poisson model.

AMS 1991 subject classifications: Primary 62–07; secondary 62F1.

ABSTRACT

The authors show that for balanced data, the estimates of effects of interest and of their standard errors are unaffected when a covariate is removed from a multiplicative Poisson model. As they point out, this is not verified in the analogous linear model, nor in the logistic model. In the first case, only the estimated coefficients remain the same, while in the second case, both the estimated effects and their standard errors can change.

RÉSUMÉ

Les auteurs montrent que dans un modèle de Poisson multiplicatif équilibré, les estimations des effets et de leur erreur standard ne sont pas affectées par le retrait d'une variable exogène. Comme ils le soulignent, ceci n'est vrai ni dans le modèle linéaire analogue, ni dans le modèle logistique. Dans le premier cas, seules les estimations des effets restent généralement les mêmes; dans le deuxième, toutes les estimations peuvent changer.

1. INTRODUCTION

Linear models, logistic models and Poisson models can be linked into a combined category called generalized linear models. Consider the case of two covariates, X and Z, and a response variable Y. Let μ denote the expected value of the response variable Y, given X and Z. The form of the generalized linear model is $g(\mu) = \beta_0 + \beta_1 X + \beta_2 Z$, where β_0, β_1 and β_2 are unknown constants. The monotone function $g(\mu)$ is called the link function.

The model is completely specified by $g(\mu)$ and a distribution for Y. In ordinary linear models, $g(\mu) = \mu$ is the identity link function, and Y is assumed to be normally distributed, while for logistic modelling, $g(\mu) = \mu/(1-\mu)$ is the logit link function, and Y has a binomial distribution. For multiplicative Poisson modelling, $g(\mu) = \log \mu$ is the log link function, and Y is a Poisson variate. In general the β's are estimated using maximum likelihood, while their standard errors are estimated from the reciprocal of the second derivative of the log-likelihood using large-sample theory (McCullagh & Nelder 1983). PROC GENMOD in SAS can be used to fit such models (PROC GENMOD 1993).

Although these three types of models are similar enough to be unified by a common theory, they behave differently with respect to the effect on estimated parameters and standard errors when a covariate is omitted. Our primary interest here is in Poisson modelling with balanced crossed data: for each combination of real-valued covariates, there are an equal number of responses. Gail (1988) has shown that removing a balanced covariate from a Poisson model results in the remaining coefficient estimate being asymptotically unbiased with an asymptotic relative efficiency of one, and in a score test comparing its asymptotically expected value with zero which has proper size. It will be shown here that the actual sample estimate of the coefficient and its standard-error estimate are the same whether the full or the reduced model is used. This will be contrasted with linear and logistic models. An example is given for demonstration of the results for the Poisson model.

2. POISSON MODELS

If a covariate is omitted from a Poisson model with balanced data, both the estimate of the remaining coefficient and the estimate of its standard error are the same as those obtained from the full model, if no offset variable (e.g., person-years) is included. This is formally stated below and proved in the appendix.

THEOREM. *Consider a set of balanced data with a levels of X, b levels of Z, and n observations, Y, for each combination of X and Z. Let Y be the dependent variable, and let X and Z be independent variables in Poisson models with no offset variables. Consider the following two models with $g(\mu) = \log \mu$:*

$$g(\mu) = \beta_0 + \beta_1 X + \beta_2 Z, \tag{1}$$

$$g(\mu) = \beta_0^* + \beta_1^* X. \tag{2}$$

Let $\hat{\beta}_1$ be the maximum-likelihood estimator of β_1 in the model (1), *and let $\hat{\beta}_1^*$ be the maximum-likelihood estimator of β_1^* in the model* (2). *Then $\hat{\beta}_1^* = \hat{\beta}_1$ and $\widehat{\text{se}}(\hat{\beta}_1^*) = \widehat{\text{se}}(\hat{\beta}_1)$.*

3. LINEAR MODELS

In linear models, if a covariate is omitted from the model, the sum of squares for that effect is pooled with the error. If the F-statistic for the removed effect was greater than 1, this results in a larger mean squared error and consequently a larger standard error of the remaining coefficients, as will be shown below. If the data are balanced, however, the estimated coefficient will be the same as that in the full model, as will also be shown.

Let the models be the same as in the theorem, except that the data are normally distributed and $g(\mu) = \mu$. Let $\mathbf{X}$ be the design matrix. In general (balanced or unbalanced data) $\hat{\beta}_1^* = \hat{\beta}_1 - c_{23}\hat{\beta}_2/c_{33}$, where c_{ij} is the (i,j)th element of $(\mathbf{X}'\mathbf{X})^{-1}$ (Snedecor & Cochran 1967). For balanced data, however, $c_{23} = 0$ and thus $\hat{\beta}_1^* = \hat{\beta}_1$. Letting c_{ij}^* be the (i,j)th element of the reduced model $(\mathbf{X}'\mathbf{X})^{-1}$, we have

$$c_{22}^* = c_{22} - c_{23}^2/c_{33} = c_{22}.$$

Thus, since $\text{Var}(\hat{\beta}_1) = \text{MSE}_f(c_{22})$ and $\text{Var}(\hat{\beta}_1^*) = \text{MSE}_r(c_{22}^*)$, the relationship between the variances of $\hat{\beta}_1$ and $\hat{\beta}_1^*$ depends solely on the relationship between the full-model mean squared error (MSE_f) and the reduced-model mean squared error (MSE_r). But

$$\text{MSE}_r = \text{MSE}_f \left\{ \frac{\text{df}\, E + (\text{df}\, Z)F}{\text{df}\, E + \text{df}\, Z} \right\},$$

where F is the F-statistic for testing $\mathcal{H}_0 : \beta_2 = 0$, and df E and df Z are the degrees of freedom for MSE_f and the mean square for Z, respectively. Thus, $\text{MSE}_r - \text{MSE}_f$ is of the same sign as $\log F$. An analogous relationship holds for the standard errors of the coefficients.

4. LOGISTIC MODELS

For logistic models, even with balanced data, the complex nature of the log likelihood function prevents the factoring of the normal equations into an X-part and a Z-part, as was done in the theorem. Thus, the estimate of the coefficient for the effect in question, as well as its estimated standard error, can be affected by the deletion of the covariate.

5. EXTENSIONS OF THE THEOREM

If the model (1) contains k covariates (with levels $a_1, \ldots, a_k$), the model (2) contains a subset of g of the k covariates, and every combination of the k covariates contains n observations, then

the conclusions of the theorem hold. The proof is notation-intensive but follows along the same lines. The scalars are replaced by vectors and matrices.

The theorem and the above extension assume that the variables involved are quantitative. If the variables are qualitative, then X in the theorem is replaced by $a-1$ zero-one dummy variables $(X_1, \ldots, X_{a-1})$ and Z is replaced by $b-1$ zero-one dummy variables $(Z_1, \ldots, Z_{b-1})$, which is similar to above extension, except that the 2^{a+b-2} combinations of the dummy variables do not occur the same number of times. For example, no combination for which $Z_1 = 1$ and $Z_2 = 1$ ever occurs. However, the $a \times b$ legitimate combinations occur together exactly n times, and thus the n factors out of the equations as it does in the proof of the theorem and its k-variable extension above. The remainder of the proof is the same as that for the k-variable extension. The following example illustrates this latter extension.

6. AN EXAMPLE

Snedecor & Cochran (1967) give an example for which the dependent variable is the number of poppy plants in oats. The data are given in Table 1.

TABLE 1: Number of poppy plants in oats.

	Treatments				
Blocks	1	2	3	4	5
1	438	538	77	17	18
2	442	422	61	31	26
3	319	377	157	87	77
4	380	315	52	16	20

The experiment is a randomized complete block design with five treatments, four blocks and one Poisson observation per treatment-block combination. Because dummy variables are used, the second extension in the previous section is appropriate. Table 2 shows the results of fitting Poisson models with SAS (PROC GENMOD 1993). The full and reduced models (deleting blocks) yield the same estimates and standard errors for each corresponding treatment.

7. DISCUSSION

Other authors have considered the asymptotic bias of coefficient estimates, the size of their standard errors, asymptotic relative efficiency, and test size when generalized linear models are employed (Gail 1986, 1988; Gail, Wieand & Piantadosi 1984; Begg & Lagakos 1993), whereas the present paper considers observed estimates of coefficients and standard errors. With data that are balanced in the sense of having an equal number of observations for each combination of the covariates, it is intuitive that the coefficient of X in the full model should be the same as that for the reduced model, because the observed effect of X cannot be due in part to Z because X and Z are uncorrelated. This in fact occurs with Poisson (with no offset variable such as person-years) and linear models, but not with logistic models. Unlike linear or logistic models, the Poisson model results in the same standard-error estimate of the covariate of interest, whether the other covariate is included in the model or not. The theorem proves this for two balanced crossed covariates, but the theorem is extendable to any number of such covariates. Such situations would only occur in

designed experiments. It is clear, however, that in such situations, omitting a covariate will maintain the properties of estimates involving the correct model, including asymptotic unbiasedness, the same efficiency, and the same size and power for Wald-based tests.

TABLE 2: PROC GENMOD results.

Parameter	Degrees of freedom	Estimate	Standard error
	Full model		
Intercept	1	3.3509	0.0901
Treatment 1	1	2.4158	0.0879
Treatment 2	1	2.4610	0.0877
Treatment 3	1	0.9006	0.0999
Treatment 4	1	0.0685	0.1171
Treatment 5	0	0.0000	0.0000
Block 1	1	0.3290	0.0469
Block 2	1	0.2265	0.0479
Block 3	1	0.2615	0.0475
Block 4	0	0.0000	0.0000
Scale	0	1.0000	0.0000
	Reduced model		
Intercept	1	3.5625	0.0842
Treatment 1	1	2.4158	0.0879
Treatment 2	1	2.4610	0.0877
Treatment 3	1	0.9006	0.0999
Treatment 4	1	0.0685	0.1171
Treatment 5	0	0.0000	0.0000
Scale	0	1.0000	0.0000

APPENDIX: PROOF OF THE THEOREM

Let Y_{ijk} be the kth value of the dependent variable for the ith level of X (X_i) and the jth level of Z (Z_j). Using the Poisson log likelihood (LL), one obtains the maximum-likelihood estimators for the model (1) as solutions to

$$\frac{\partial \mathrm{LL}}{\partial \tilde{\beta}} = \begin{bmatrix} \sum\sum\sum Y_{ijk} - \sum\sum\sum E_{ij} \\ \sum\sum\sum X_i Y_{ijk} - \sum\sum\sum X_i E_{ij} \\ \sum\sum\sum Z_j Y_{ijk} - \sum\sum\sum Z_j E_{ij} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

where $E_{ij} = \exp(\beta_0 + \beta_1 X_i + \beta_2 Z_j)$ and $\tilde{\beta}$ is the vector containing the three betas. The first equation can be rewritten as

$$\begin{aligned} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} E_{ij} = n\left(\sum_{i=1}^{a} e^{\beta_0+\beta_1 X_i}\right)\left(\sum_{j=1}^{b} e^{\beta_2 Z_j}\right) \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} D e^{\beta_0+\beta_1 X_i}/b = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} e^{\gamma_0+\beta_1 X_i}, \end{aligned}$$

where $D = \sum_{j=1}^{b} e^{\beta_2 Z_j}$ and $\gamma_0 = \beta_0 + \log(D/b)$. The second equation can also be rewritten as

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_i Y_{ijk} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_i e^{\gamma_0+\beta_1 X_i}.$$

Hence if β_2 is considered fixed, the first two equations of $\partial \mathrm{LL}/\partial\tilde{\beta} = 0$ for the model (1) have the same form as the corresponding first two equations for the model (2). Therefore, $\hat{\beta}_1^* = \hat{\beta}_1$ and $\hat{\beta}_0^* = \hat{\beta}_0 + \log(\hat{D}/b) = \hat{\beta}_0 + \log(\sum_j e^{\hat{\beta}_2 Z_j}/b)$.

The variance estimators are obtained from the negative reciprocal of the second derivative of the log likelihood evaluated at the maximum-likelihood estimator. Note that

$$-\frac{\partial^2 \mathrm{LL}}{\partial \tilde{\beta}^2} = \begin{bmatrix} \sum\sum\sum E_{ij} & \sum\sum\sum X_i E_{ij} & \sum\sum\sum Z_j E_{ij} \\ \sum\sum\sum X_i E_{ij} & \sum\sum\sum X_i^2 E_{ij} & \sum\sum\sum X_i Z_j E_{ij} \\ \sum\sum\sum Z_j E_{ij} & \sum\sum\sum X_i Z_j E_{ij} & \sum\sum\sum Z_j^2 E_{ij} \end{bmatrix}.$$

Using the cofactor form of $(-\partial^2 \mathrm{LL}/\partial\tilde{\beta}^2)^{-1}$, one obtains that its $(2,3)$ and $(3,2)$ entries are 0, because

$$\begin{aligned} &\left(\textstyle\sum_i \sum_j \sum_k E_{ij}\right)\left(\sum_i \sum_j \sum_k X_i Z_j E_{ij}\right) \\ &= \left\{n\left(\sum_i e^{\beta_0+\beta_1 X_i}\right)\left(\sum_j e^{\beta_2 Z_j}\right)\right\}\left\{n\left(\sum_i X_i e^{\beta_0+\beta_1 X_i}\right)\left(\sum_j Z_j e^{\beta_2 Z_j}\right)\right\} \\ &= \left(\sum_i\sum_j\sum_k X_i E_{ij}\right)\left(\sum_i\sum_j\sum_k Z_j E_{ij}\right). \end{aligned}$$

Using the notation of Searle (1971), write

$$-\frac{\partial^2 \mathrm{LL}}{\partial \tilde{\beta}^2} = \begin{bmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{bmatrix} \quad \text{and} \quad \left(-\frac{\partial^2 \mathrm{LL}}{\partial \tilde{\beta}^2}\right)^{-1} = \begin{bmatrix} W_{11} & W_{12} \\ W_{12}' & W_{22} \end{bmatrix},$$

where V_{11} is the upper left-hand 2×2 submatrix of $-\partial^2 \mathrm{LL}/\partial \tilde{\beta}^2$. Then $V_{11}^{-1} = W_{11} - W_{12}W_{22}^{-1}W_{12}'$. Since the $(3,2)$ element of $-(\partial^2 \mathrm{LL}/\partial \tilde{\beta}^2)^{-1}$ is equal to 0, $W_{12}' = [c_{13} \quad 0]$, and hence

$$V_{11}^{-1} = W_{11} - \begin{bmatrix} c_{31} \\ 0 \end{bmatrix} \frac{1}{c_{33}} [c_{13} \quad 0] = W_{11} - \begin{bmatrix} c_{31}c_{13}/c_{33} & 0 \\ 0 & 0 \end{bmatrix},$$

where c_{ij} is the (i,j)th element of $(-\partial^2 \mathrm{LL}/\partial \tilde{\beta}^2)^{-1}$. Hence the $(2,2)$ element of W_{11}, which is also the $(2,2)$ element of $(-\partial^2 \mathrm{LL}/\partial \tilde{\beta}^2)^{-1}$, is equal to the $(2,2)$ element of V_{11}^{-1}. Thus

$$\begin{aligned} \widehat{\mathrm{Var}}(\hat{\beta}_1) &= \left[(2,2) \text{ entry of } \left(-\frac{\partial^2 \mathrm{LL}}{\partial \tilde{\beta}^2}\right)^{-1} \text{ for model (1)}\right]_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \\ &= \left[(2,2) \text{ element of } V_{11}^{-1}\right]_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} = \left[\frac{(1,1) \text{ element of } V_{11}}{\det(V_{11})}\right]_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \end{aligned}$$

and hence $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ is equal to

$$\frac{\sum\sum\sum e^{\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_j}}{\left(\sum\sum\sum e^{\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_j}\right)\left(\sum\sum\sum X_i^2 e^{\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_j}\right) - \left(\sum\sum\sum X_i e^{\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_j}\right)^2}.$$

Note that

$$\begin{aligned} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} e^{\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_j} &= \sum_{i=1}^{a}\sum_{k=1}^{n} e^{\hat{\beta}_1 X_i} b e^{\hat{\beta}_0 \left(\sum_{j=1}^{b} e^{\hat{\beta}_2 Z_j}/b\right)} \\ &= \sum_{i=1}^{a}\sum_{k=1}^{n} e^{\hat{\beta}_1 X_i} b e^{\hat{\beta}_0 + \log\left(\sum_{j=1}^{b} e^{\hat{\beta}_2 Z_j}/b\right)} \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} e^{\hat{\beta}_0^* + \hat{\beta}_1^* X_i}. \end{aligned}$$

Similar results hold for the analogous summations containing X_i factors. Thus

$$\begin{aligned} \widehat{\mathrm{Var}}(\hat{\beta}_1) &= \frac{\sum\sum\sum e^{\hat{\beta}_0^* + \hat{\beta}_1^* X_i}}{\left(\sum\sum\sum e^{\hat{\beta}_0^* + \hat{\beta}_1^* X_i}\right)\left(\sum\sum\sum X_i^2 e^{\hat{\beta}_0^* + \hat{\beta}_1^* X_i}\right) - \left(\sum\sum\sum X_i e^{\hat{\beta}_0^* + \hat{\beta}_1^* X_i}\right)^2} \\ &= \left[(2,2) \text{ entry of } \left(-\frac{\partial^2 \mathrm{LL}}{\partial \tilde{\beta}^2}\right)^{-1} \text{ for model (2)}\right]_{\hat{\beta}_0^*, \hat{\beta}_1^*} = \widehat{\mathrm{Var}}(\hat{\beta}_1^*). \end{aligned}$$

This completes the proof. □

REFERENCES

M. D. Begg & S. Lagakos (1993). Loss in efficiency caused by omitting covariates and misspecifying exposure in logistic regression models. *Journal of the American Statistical Association*, 88, 166–170.

M. H. Gail (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology* (S. H. Moolgavkar & R. L. Prentice, eds.), Wiley, New York, pp. 3–18.

M. H. Gail (1988). The effect of pooling across strata in perfectly balanced studies. *Biometrics*, 44, 151–162.

M. H. Gail, S. Wieand & S. Piantadosi (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71, 431–444.

P. McCullagh & J. A. Nelder (1983). *Generalized Linear Models*. Chapman & Hall, New York.

PROC GENMOD (1993). SAS Technical Report P-243, SAS Institute, Cary, NC.

S. R. Searle (1971). *Linear Models*. Wiley, New York.

G. W. Snedecor & W. G. Cochran (1967). *Statistical Methods*. The Iowa State University Press, Ames, IA.

Received 1 July 1998
Accepted 27 September 1999

Martin R. PETERSEN: mrp1@cdc.gov
James A. DEDDENS: jad0@cdc.gov
National Institute for Occupational Safety and Health, Mail Stop R13
4676 Columbia Parkway, Cincinnati, OH 45226-1998, USA