American Society for Quality

Adding a Variable in Generalized Linear Models
Author(s): P. C. Wang
Source: *Technometrics*, Vol. 27, No. 3 (Aug., 1985), pp. 273-276
Published by: American Statistical Association and American Society for Quality
Stable URL: http://www.jstor.org/stable/1269708
Accessed: 01/05/2014 22:27

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association and *American Society for Quality* are collaborating with JSTOR to digitize, preserve and extend access to *Technometrics.*

http://www.jstor.org

This content downloaded from 140.113.38.11 on Thu, 1 May 2014 22:28:00 PM
All use subject to JSTOR Terms and Conditions

© 1985 American Statistical Association and
the American Society for Quality Control

TECHNOMETRICS, AUGUST 1985, VOL. 27, NO. 3

Adding a Variable in Generalized Linear Models

P. C. Wang

Department of Applied Mathematics
National Chiao Tung University
Hsinchu, Taiwan
Republic of China

The likelihood ratio statistic can be used to determine the significance of an explanatory variable in a generalized linear model. In order to obtain such a statistic, however, we need two sets of iterations for two maximum likelihoods. Moreover, the statistic is not directed to detect influential or outlying observations that affect the importance of the variable considered. Therefore we develop relatively simple procedures to help the analyst select an appropriate model and detect the effects of such observations on adding a variable into any model. Two examples are given for illustration.

KEY WORDS: Added variable plots; Score statistics; Influence; Outliers; GLIM.

1. INTRODUCTION

Diagnostics are used to assess the adequacy of assumptions underlying a model and to identify unexpected characteristics of the data that may seriously influence conclusions or require special attention. The importance of diagnostics in normal regression analysis has been emphasized by several authors (e.g., Belsley et al. 1980, Cook and Weisberg 1982, and Hocking 1983). A variety of graphical and nongraphical methods for this purpose are available to aid in an analysis based on the linear model with normal errors, but relatively few methods have been developed for application outside this framework. Pregibon (1981), Landwehr et al. (1984), and Cook and Wang (1983) are three exceptions. The first two develop diagnostic procedures for logistic regression and the last do this for nonlinear regression.

Many diagnostic procedures are available for selecting explanatory variables and detecting the effects of influential or outlying observations on a particular explanatory variable under a normal linear model. One of them, the added variable plot, serves well for this purpose, as illustrated in Cook and Weisberg (1982). Atkinson (1982, 1983) and Cook and Wang (1983) developed similar plots for transformations and nonlinear regression. In this article I extend these diagnostic approaches to generalized linear models that have been applied to data in many fields, as illustrated by McCullagh and Nelder (1983).

The technique of the score statistic has been used to derive diagnostic procedures for various purposes. Atkinson (1982) used it to construct partial residual plots for diagnosing the need for transformation of the responses in linear regression; Cook and Weisberg (1983) used it and its graphical version for diagnostic information about heteroscedasticity; Pregibon (1982) and Chen (1983) used it to test the need for explanatory variables in generalized linear models. The main exploration here is to use the score statistic and its graphical version to determine if an explanatory variable is significant. I derive results in the next section and present two applications in Section 3. Concluding comments are given in the last section.

2. THE SCORE STATISTIC AND ADDED VARIABLE PLOT

The probability function for a single response y in a generalized linear model (McCullagh and Nelder 1983) is

$$h(y \mid \theta, \phi) = \exp\{[\theta y - b(\theta)]/a(\phi) + c(y, \phi)\}$$

for some smooth functions $a(\ )$, $b(\ )$, and $c(\ )$. Our concern is the parameter θ. Let $(\beta^T, \gamma)^T$ be a vector of $p+1$ parameters, $(x^T, z)^T$ be a vector of observed values of $p+1$ explanatory variables, and $\eta = x^T\beta + z\gamma$ be a link function of $E(y) = \dot{b}(\theta)$, the derivative of the function $b(\theta)$. Our main interest is to use the score statistic, instead of the likelihood ratio test (LRT) statistic, to determine if $\gamma = 0$. Given a sample of independent observations (y_i, x_i^T, z_i), assume $\theta_i = g(\eta_i) = g(x_i^T\beta + z_i\gamma)$ for a fixed function g. The log-likelihood function of (β^T, γ) for the sample is

$$L(\theta, Y) = \sum_{i=1}^{n} \{[\theta_i y_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}, \quad (1)$$

where $Y = (y_1, \ldots, y_n)^T$.

Under the null hypothesis H_0 that $\gamma = 0$, the maxi-

This content downloaded from 140.113.38.11 on Thu, 1 May 2014 22:28:00 PM
All use subject to JSTOR Terms and Conditions

mum likelihood estimate (MLE) $\hat{\beta}$ of β can be obtained by an iterative method. Let $\cdot$ denote the differentiation sign. We obtain the score for H_0,

$$U(\hat{\beta}) = \sum_{i=1}^{n} \dot{g}(\hat{\eta}_i)[y_i - \dot{b}(\hat{\theta}_i)]z_i/a(\phi), \tag{2}$$

and the information matrix for (β^T, γ),

$$I(\beta, \gamma) = \sum_{i=1}^{n} d(\theta_i)\begin{pmatrix} x_i x_i^T & x_i z_i \\ z_i x_i^T & z_i z_i \end{pmatrix}, \tag{3}$$

where $\hat{\theta}_i = g(\hat{\eta}_i)$, $\hat{\eta}_i = x_i^T\hat{\beta}$, and $d(\theta_i) = \dot{g}(\eta_i)^2\ddot{b}(\theta_i)/a(\phi)$. Let $X = (x_1, \ldots, x_n)^T$, $Z = (z_1, \ldots, z_n)^T$, $V = \text{diag } \{d(\hat{\theta}_i)\}$, and S be a vector of entries $\dot{g}(\hat{\eta}_i)[y_i - \dot{b}(\hat{\theta}_i)]/a(\phi)$. Then (2) and (3) can be written in simpler forms:

$$U(\hat{\beta}) = S^TZ$$

and

$$I(\hat{\theta}) = \begin{pmatrix} X^TVX & X^TVZ \\ Z^TVX & Z^TVZ \end{pmatrix}.$$

After a few algebraic manipulations, the score statistic for H_0 (Cox and Hinkley 1974, p. 324), $U(\hat{\beta})^T I_z^{-1} U(\hat{\beta})$, becomes

$$T = (S^TZ)^2/(V^{1/2}Z)^T(I - H)V^{1/2}Z, \tag{4}$$

where $H = V^{1/2}X(X^TVX)^{-1}X^TV^{1/2}$ and I_z^{-1} is the $((p + 1), (p + 1))$ entry of the inverse of $I(\hat{\theta})$. This statistic is computationally equivalent to the F test for adding a variable in a normal linear model. Using this fact, we can establish a graphical version of (4) for diagnostic purposes.

Let $W = V^{1/2}X\hat{\beta} + V^{-1/2}S$. We construct the approximating model,

$$W = V^{1/2}X\beta + V^{1/2}Z\gamma + E, \tag{5}$$

where E is $N(0, \sigma^2 I)$ distributed. The added variable plot for variable $V^{1/2}Z$ under (5) is a plot of $R = V^{-1/2}S$ versus $(I - H)V^{1/2}Z$ (Cook and Weisberg 1982). I use this plot for diagnostic checks, since the significance of the F test for $\gamma = 0$ under model (5), which is equivalent to (4), corresponds to the "nonzeroness" of the slope of the regression line in the plot. Note that the regression line passes through the origin because we can always find a vector perpendicular to both R and $(I - H)V^{1/2}Z$.

For convenience, the entries of R and $(I - H)V^{1/2}Z$ are called residuals and Z-residuals, respectively. Generally we might need extra computations to obtain V. However, under the most commonly used models, such as normal, binomial, Poisson, and exponential models, it is available after the iteration to obtain the MLE of β.

In the rest of the section, I explore our discussions further for several special distributions.

Case 1. Let $b(\theta) = \theta^2/2$, $a(\phi) = \sigma^2$, and g be the identity function. This is a normal linear model—that is, $Y = X\beta + Z\gamma + E$ with $E \sim N(0, \sigma^2 I)$, $V = \sigma^{-2}I$, and $R = (Y - X\hat{\beta})/\sigma$ when σ is known. When σ^2 is unknown, we need to replace σ^2 by its MLE $\hat{\sigma}^2$ to compute V, R, and the score statistic. This score statistic is equivalent to the F-test statistic. The plot that I propose is equivalent to the added variable plot illustrated in section 2.3.2 of Cook and Weisberg (1982). The usefulness of such plots is also discussed in the book, with several numerical examples.

Case 2. g is the identity, and $a(\phi)$ is constant. The logistic regression, the Poisson regression with mean e^{η}, and the exponential regression with mean η^{-1} are three examples. Denote $V^* = \text{diag } \{d(\theta_i) - \ddot{g}(\eta_i)(y_i - \dot{b}(\theta_i))/a(\phi)\}$. The Newton–Raphson method for finding the MLE of β leads to the iterative scheme

$$\beta^{t+1} = \beta^t + (X^TV^*X)^{-1}X^TS, \qquad t = 0, 1, 2, \ldots, \tag{6}$$

where V^* and S are evaluated at β^t. At the "final" step, $V = V^* = \text{diag } \{d(\hat{\theta}_i)\}$ is available with $\hat{\beta}$. In fact, the information matrix in this case is equal to the negative of the second derivative matrix of the log-likelihood; that is not necessarily true in general.

Case 3. The general case: The iterative scheme (6) can be applied with the same V^*. However, we need to recalculate V after obtaining the MLE of β. In some cases, V becomes the identity matrix (e.g., for the exponential distribution with mean e^{η}).

3. EXAMPLES

In this section I present two brief examples to illustrate the usefulness of added variable plots in generalized linear models, one in logistic regression and the other in Poisson regression. I omit those in normal linear regression, since they have been illustrated by Cook and Weisberg (1982).

3.1 Finney's Data

The data were analyzed by Finney (1947). Pregibon (1981) used them to illustrate the building blocks of logistic regression diagnostics. The response is either the occurrence or nonoccurrence of vasoconstriction. This response might depend on the rate (X_1) and volume (X_2) of air inspired on a transient vasoconstriction in the skin of the digits. The plots in Figures 1 and 2, where the score statistics' values are 14.82 and 13.68, respectively, clearly indicate the significance of one variable after the other is included in the model. This suggests the importance of these two explanatory variables. Note that observations 4 and 18 in both plots seem to be the two most important observations that affect the slope of the regression through the origin, or the significance of the score statistic. In fact, when these two observations are deleted, the MLE of β changes from $(-2.875, 5.719, 4.526)$ to $(-24.58, 39.55, 31.94)$. Observation 32,

This content downloaded from 140.113.38.11 on Thu, 1 May 2014 22:28:00 PM
All use subject to JSTOR Terms and Conditions

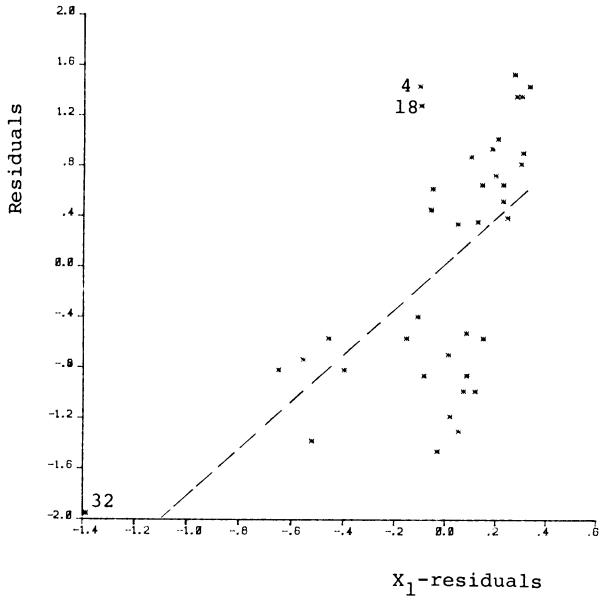

Figure 1. Added Variable Plot for Rate in Finney's (1947) Data.

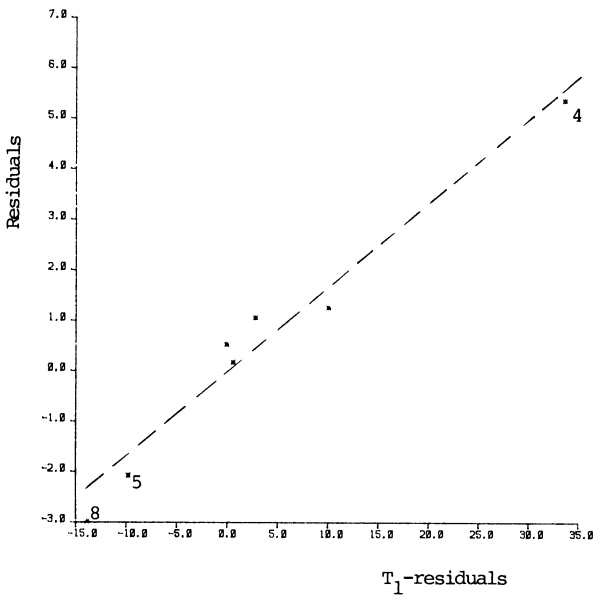

Figure 3. Added Variable Plot for Time T_1 in Jørgensen's (1961) Data.

which stands out in Figure 1, also attracts our attention, although it does not affect the regression much. Its strange X_1 value can be detected after careful checks.

3.2 Jørgensen's Data

The data are taken from Jørgensen (1961). The response is the number of failures of a complex piece of electronic equipment in a week. Nine weeks in which each week was divided into two operating regimes were spent in observation. Explanatory variables T_1 and T_2 are times spent in regimes one and two, respectively. For my illustration, treat these nine responses as independent Poisson observations with possible mean $T_1\beta_1 + T_2\beta_2$. The importance of explanatory variable T_1 after including T_2 in the model is clearly shown in the plot of Figure 3 with score statistic 39.99. Moreover, the coefficient of T_1 estimated by the plot is .159, not far from its MLE of .167. Unlike time T_1, however, time T_2 is not significant when time T_1 is included. The added variable plot in Figure 4 indicates this with a value of 1.825 for the score statistic.

Observations 4, 5, and 8 seem to control two linear

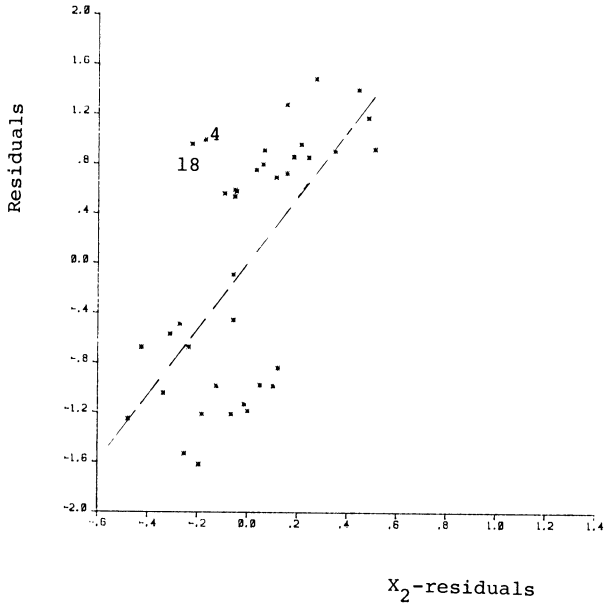

Figure 2. Added Variable Plot for Volume in Finney's (1947) Data.

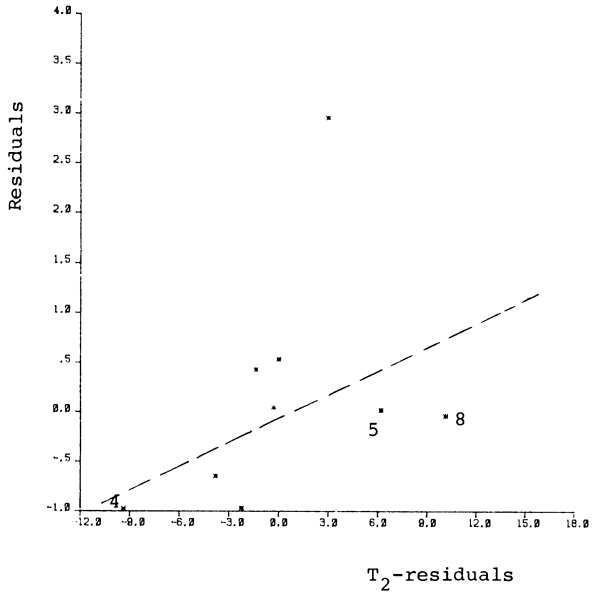

Figure 4. Added Variable Plot for Time T_2 in Jørgensen's (1961) Data.

This content downloaded from 140.113.38.11 on Thu, 1 May 2014 22:28:00 PM
All use subject to JSTOR Terms and Conditions

trends on these two added variable plots. Both plots suggest that the conclusion would be totally different without these three observations. Moreover, observation 1 in Figure 4, which has the largest residual, suggests the decrease of the importance of variable T_2 without it. These suggestions are confirmed by the score statistics without the corresponding observations.

4. DISCUSSIONS AND CONCLUSIONS

Pregibon (1981) gave excellent diagnostic procedures in logistic regression, based on the model that includes all of the important explanatory variables. In the preceding, I provide another approach to better understanding of the nature of the data under consideration. My procedures, both the score statistic and the corresponding added variable plot, check the importance of any explanatory variable and possibly influential or outlying observations for this variable. The score statistic indicates the significance of the slope of the regression line in the added variable plot and determines the importance of the underlying variable. The plot gives the overall impression of the scatter of the data points and detects observations that affect the importance of the variable. Such observations might be revealed when they are separated from the rest of the data. Thus I view these two procedures as complementary, and I recommend the use of both.

Both the score statistic and the LRT statistic can be used for generalized linear models and be computed in the GLIM computer package. However, score statistics need fewer computations, since no iteration to obtain MLE's is needed under the alternative. Although the score statistic I derived is a special case of Pregibon's (1982), we expect it to be more useful along with its corresponding added variable plot.

The slope of the regression line in an added variable plot can be used as an estimate of the coefficient of the variable considered. In normal linear regression, this slope is exactly its MLE under the alternative. However, this is not true in general. The MLE's in generalized linear models are usually obtained by an iterative method. Thus it is almost impossible to obtain such an estimate in any single step. The slope of the regression line in the plot might be a good starting estimate of the coefficient considered.

ACKNOWLEDGMENT

I thank the referees, the associate editor, and the editor for their helpful comments that led to improvements in the presentation of this article.

[*Received February 1984. Revised October 1984.*]

REFERENCES

Atkinson, A. C. (1982), "Regression Diagnostics, Transformations and Constructed Variables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 44, 1–36.

——— (1983), "Diagnostics Regression Analysis and Shifted Power Transformations," *Technometrics*, 25, 23–34.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley.

Chen, C. (1983), "Score Tests for Regression Models," *Journal of the American Statistical Association*, 78, 158–161.

Cook, R. D., and Wang, P. C. (1983), "Diagnostics for Nonlinear Regression," unpublished manuscript.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

——— (1983), "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, 70, 1–10.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Finney, D. J. (1947), "The Estimation From Individual Records of the Relationships Between Dose and Quantal Response," *Biometrika*, 34, 320–324.

Hocking, R. R. (1983), "Developments in Linear Regression Methodology: 1959–1982" (with discussion), *Technometrics*, 25, 219–249.

Jørgensen, D. W. (1961), "Multiple Regression Analysis of a Poisson Process," *Journal of the American Statistical Association*, 56, 235–245.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984) "Graphical Methods for Assessing Logistic Regression Models" (with discussion), *Journal of the American Statistical Association*, 79, 61–83.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman and Hall.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *Annals of Statistics*, 9, 705–724.

——— (1982), "Score Tests in GLIM With Applications," in *GLIM.82: Proceedings of the International Conference on Generalized Linear Models* (Ser. 14, "Lecture Notes in Statistics"), New York: Springer-Verlag.

This content downloaded from 140.113.38.11 on Thu, 1 May 2014 22:28:00 PM
All use subject to JSTOR Terms and Conditions