

July 31, 1996

Generalized Linear Models: Checking Assumptions and Strengthening Conclusions

by

N. E. Breslow

Department of Biostatistics
University of Washington
Seattle WA 98195-7232, USA

Prepared for the
Congresso Nazionale
Societa' Italiana di Biometria
Centro Convegni S. Agostino, Cortona
16-17 June, 1995

ABSTRACT

Key assumptions that underlie the application of standard generalized linear models (GLMs) include the statistical independence of the observations, the correct specification of the link and variance functions, the correct scale for measurement of the explanatory variables and the lack of undue influence of individual observations on the fitted model. Using data on counts of epileptic seizures before and after treatment (Thall and Vail, 1990) for illustration, this paper reviews methods that may be applied after fitting a GLM to detect departures from the key assumptions and to make valid inferences even in the face of them. Problems of overdispersion may be resolved by recourse to the jackknife, the bootstrap or the “sandwich” standard error estimates, for example, as well as by fitting of models with parameters in the variance function. Many of the techniques are easily implemented in the S statistical language by calling routines already developed for linear model analysis.

1. INTRODUCTION

The generalized linear model (GLM) [1] neatly synthesizes many of the most commonly used statistical techniques for the analysis of both continuous and discrete data in a unified conceptual and methodological framework. It permits the adaptation of procedures for model building and model checking, originally developed for normal theory linear regression, for use in a much wider setting. Many of these techniques have been incorporated into the the S statistical language [2] and in macros or “S-functions” written by statisticians in order to implement their enhancements of GLMs. This paper reviews these developments and illustrates their use in practice by application to a problem where the response measurements are counts that might be considered ordinarily to have a Poisson distribution.

Suppose one observes data (y_i, x_i, w_i) , $i = 1, \dots, n$, for a series of n subjects where y_i denotes a univariate response measurement, x_i is a p -vector of explanatory variables and w_i is a *prior weight* that specifies the precision of y_i . When y is a binomial proportion, for example, w is the denominator. A GLM is specified by two key functions g and v . The *link function* $g(\mu) = \eta$ relates the mean $\mu = E(y)$ to the *linear predictor* $\eta = x\beta$, where β is a p -vector of regression coefficients. The *variance function* v relates the variance to the mean by $\text{Var}(y) = \phi w^{-1}v(\mu)$, where ϕ is a dispersion factor. For the Poisson and binomial models, where $v(\mu) = \mu$ and $v(\mu) = \mu(1 - \mu)$, respectively, $\phi = 1$ and the variance is completely specified by the mean. For the normal theory linear model μ is the identity and v is the constant 1, but the unknown variance $\sigma^2 = \phi$ is generally estimated from the data. The two functions g and v thus define how the GLM generalizes ordinary least squares regression. The Appendix contains a brief review of GLM theory and nomenclature.

The critical assumptions that underlie the GLM, many of which apply to any regression model, are the following:

- Statistical independence of the n observations
- Correct specification of the variance function v
- Correct specification of ϕ (1 for Poisson and binomial data)
- Correct specification of the link function g
- Correct form for the explanatory variables x
- Lack of undue influence of individual observations on the fit

Failure of any one of these assumptions may seriously compromise the conclusions of the data analysis. This paper presents methods to check these assumptions and describes extensions of standard GLM methodology that help to guard against reaching false conclusions when certain assumptions fail to hold.

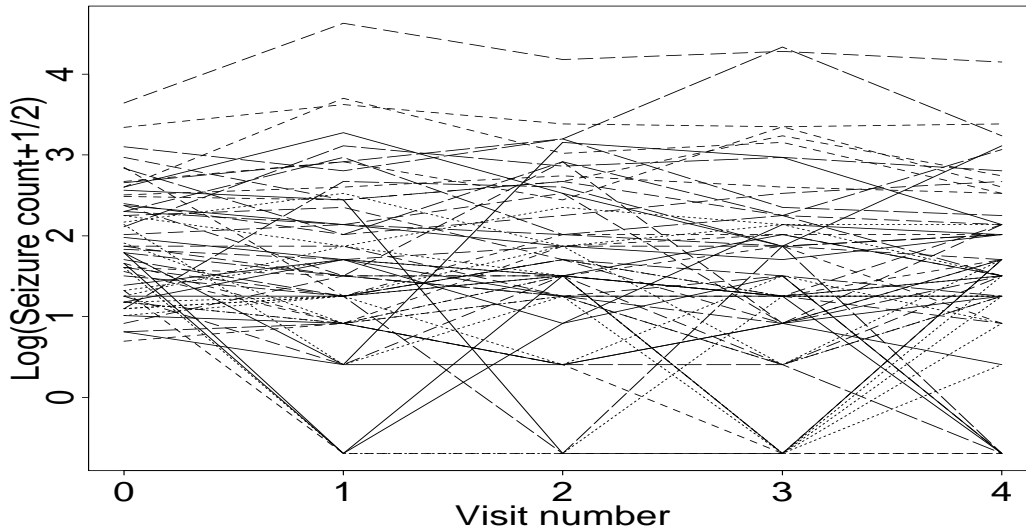


Figure 1: Log transforms of epilepsy seizure counts at baseline and at four follow-up periods.

2. THE EPILEPSY DATA

Thall and Vail [4] reported data from a clinical trial of 59 patients with epilepsy, 31 of whom were randomized to receive the anti-epilepsy drug Progabide and 28 of whom received a placebo. Baseline data consisted of the patient's age and the number of epileptic seizures recorded during an 8 week period prior to randomization. The response consisted of counts of seizures occurring during the two week period prior to each of four follow-up visits. Figure 1 graphs the logarithms of the bi-weekly seizure rates reported at baseline and at each of the four visits. Counts for the latter were increased by $1/2$ in order to avoid infinities when taking logs. The variability of the baseline seizure rate is less because the period over which it was measured was four times as long as for each of the post treatment measurements. One patient, #207, had an exceptionally high baseline count that *increased* following treatment with Progabide.

Figure 2 shows the mean responses (log seizure rate) at baseline and at each subsequent visit. Progabide appears to have reduced the seizure rate slightly, with the responses at the second visit being an exception. Since this paper concerns the GLM with *univariate* response, the total seizure count over the four follow-up periods is used as the response variable. (An exception is made in §6 where we treat the four follow-up counts as replicate measurements having the same mean structure in order to briefly consider issues posed by dependent data.) Conventional univariate analyses include each individual's baseline count as a covariable in the model equation [4]. Thus the statistical dependence between pre- and post- randomization seizure

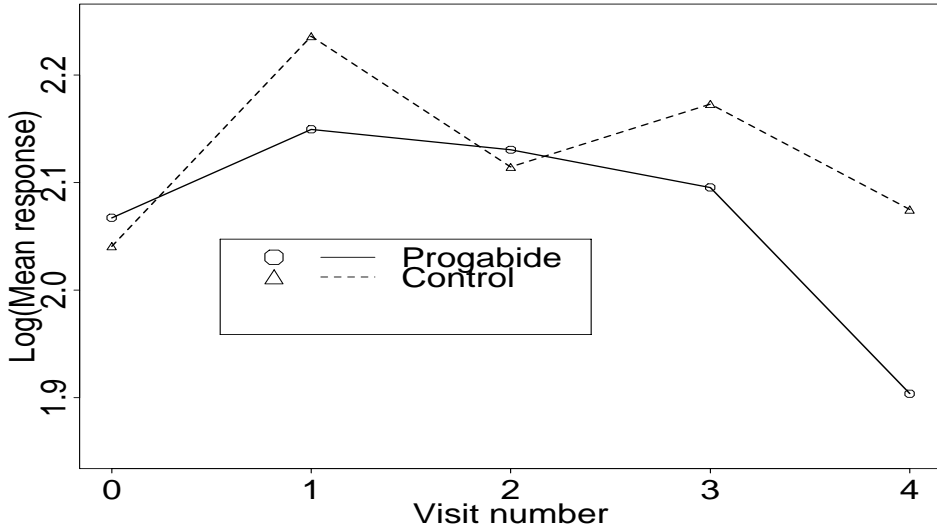


Figure 2: Log transforms of mean values of epilepsy seizure counts for treated and control subjects.

counts for the same individual is accounted for in the mean structure of the model. Regression coefficients for the treatment effect and treatment \times baseline interaction model the average treatment effect, considered as a function of the baseline count. When measured as logarithm of the ratio of the treatment to control seizure rates, for example, the treatment effect appears more pronounced among subjects who had lower counts to begin with (Figure 3). Alternative *longitudinal* analyses of these same data, by contrast, account for the within subject dependence between baseline and subsequent counts in the correlation structure of the model [6]. The treatment effect is expressed as the cross products ratio from the 2×2 table of mean seizure rates: pre- vs. post-, and treatment vs. control. It thus measures the multiplicative effect of treatment on the ratio of population means where one averages over both pre- and post- randomization counts.

3. DELETION DIAGNOSTICS

The influence of individual observations on a GLM fit is measured by the change in the estimated regression coefficients that result from their removal from the dataset. Thus we study $\hat{\beta}_{(i)} - \hat{\beta}$ for all i where $\hat{\beta}$ is the estimate based on all n subjects and $\hat{\beta}_{(i)}$ is the estimate when the i^{th} observation is deleted. This is computationally burdensome because of the need to iterate the fitting procedure to convergence with n new sets of data. By applying standard updating formulae for least squares regression to the GLM *working residuals*, however, we obtain the results of the first iteration of the Fisher scoring algorithm applied to each of the n reduced datasets starting

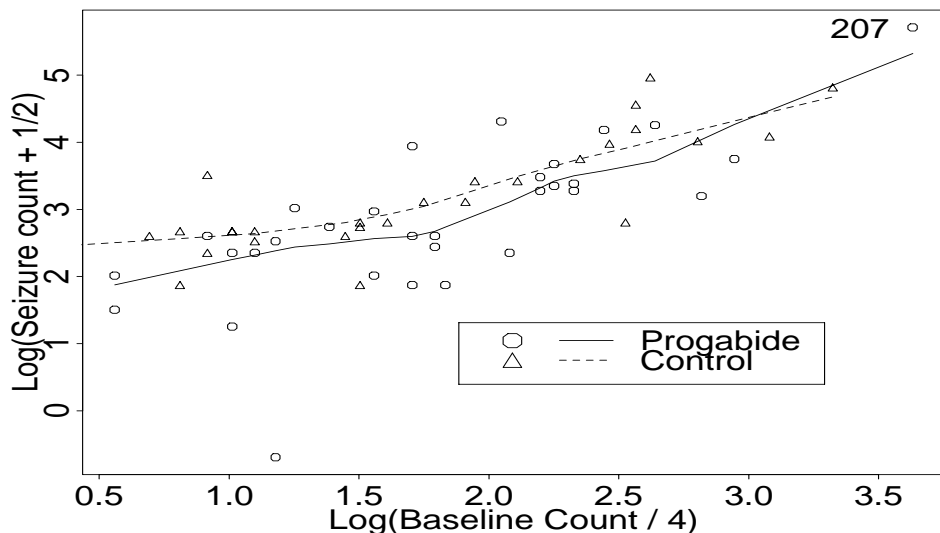


Figure 3: Scatterplot of log baseline counts against log follow-up count for treated and control subjects, with *loess* smooths.

at $\hat{\beta}$ [5]. These “one-step” approximations $\hat{\beta}_{(i)}^1$ are sufficiently close to $\hat{\beta}_{(i)}$ for most practical purposes. Normalizing the differences by the Fisher information \mathcal{I} , *i.e.*, by the inverse asymptotic variance, we obtain a measure $c_i = (\hat{\beta}_{(i)}^1 - \hat{\beta})^t \mathcal{I} (\hat{\beta}_{(i)}^1 - \hat{\beta})$ known as Cook’s distance that roughly calibrates the difference between $\hat{\beta}_{(i)}^1$ and $\hat{\beta}$ in units of the χ_p^2 distribution [5].

With y_i denoting the sum of the seizure counts during the four followup periods, we first fit a log-linear Poisson regression model using as explanatory variables the logarithm of 1/4 the number of seizures during the baseline period, the logarithm of age, a binary indicator of treatment and the interaction (product) of treatment with log baseline count. Figure 4 shows excellent agreement between the true and one-step deletion diagnostics for the treatment coefficient and indicates that subject #207 has an enormous influence on the fit. This is confirmed by the Cook’s distances (Figure 5). Inclusion of #207 in the analysis substantially alters both the intercept and slope of the fitted regression relationship for treated patients (Figure 6). Such instability was not evident from the normal quantile plot of *deviance residuals* (Figure 7). Subject #207 was removed from all further analyses of these data.

4. CHOOSING THE RIGHT TRANSFORMATION

A reasonable strategy for the analysis of clinical trial data is to first consider the effects of the covariables on the response, modelling them in a reasonably flexible way, and then to consider the treatment effects and the treatment by covariable interactions. A question that arises is how best to model the covariable effects. In least squares

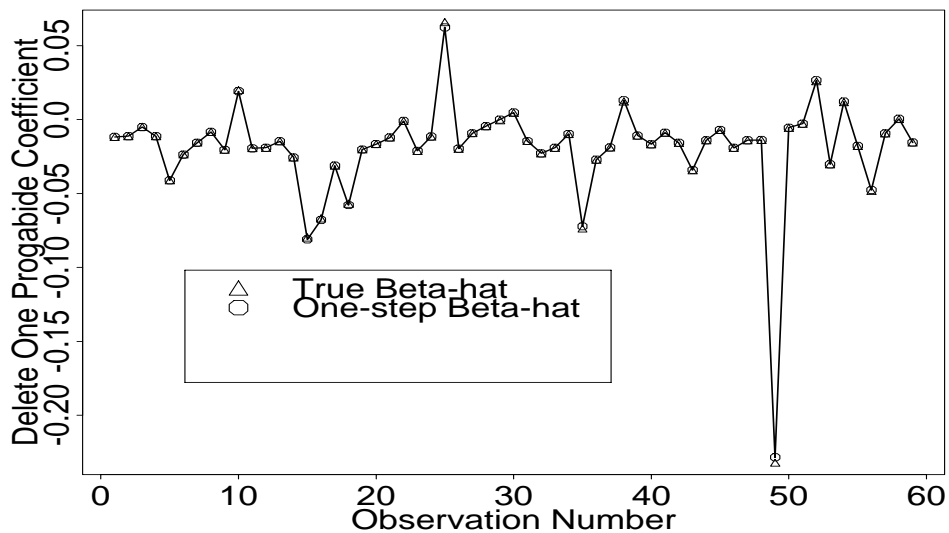


Figure 4: Effect of deletion of each subject on the regression coefficient for Progabide, and the one-step approximations.

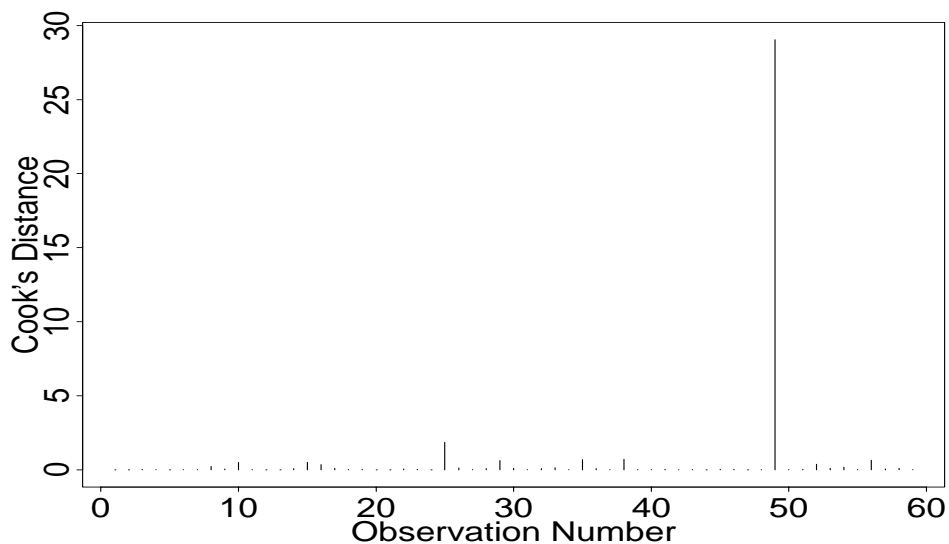


Figure 5: Cook's distance measure of leverage.

regression, scatterplots of *partial residuals* of the form $y_i - \hat{\mu}_i + x_i^{(k)} \hat{\beta}_k$ against each of the continuous covariables $x_i^{(k)}$ are used to determine whether some transformation of $x^{(k)}$ may be needed in order to achieve a linear relationship [5]. The same technique, when applied with the GLM *working residuals* in place of $y_i - \hat{\mu}_i$, is valuable for GLMs. A smoother such as *loess* is indispensable in order to extract the relevant

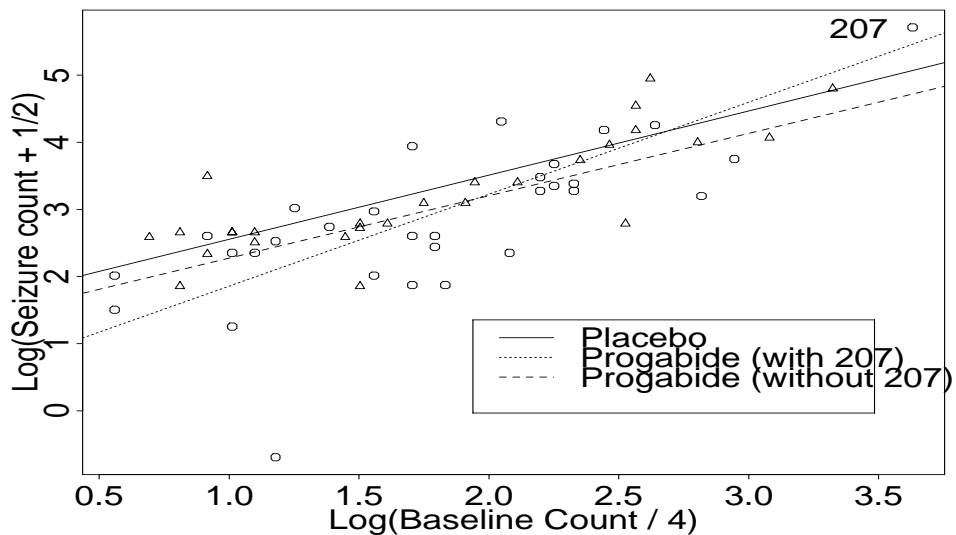


Figure 6: Change in fitted regression lines from deletion of subject #207.

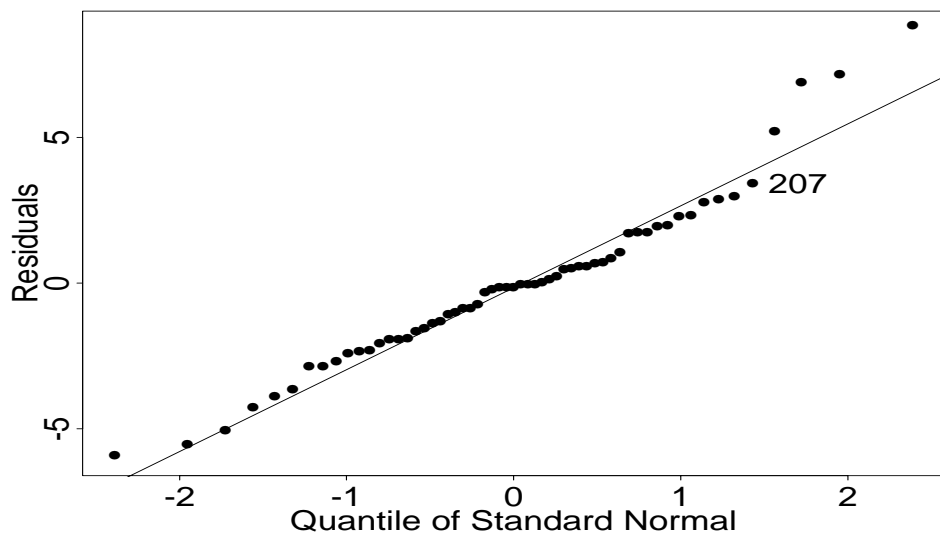


Figure 7: Normal quantile plot of deviance residuals.

information from the scatterplot [2].

To illustrate this technique, we fit a Poisson regression model to the epilepsy data with linear terms in age and baseline count as the only covariables. Partial residual plots (Figures 8 and 9) confirm that our original selection of a log transform for the baseline count was wise, but that a simple linear term in age could suffice. Test statistics described in the next section indicated the presence of a positive interaction

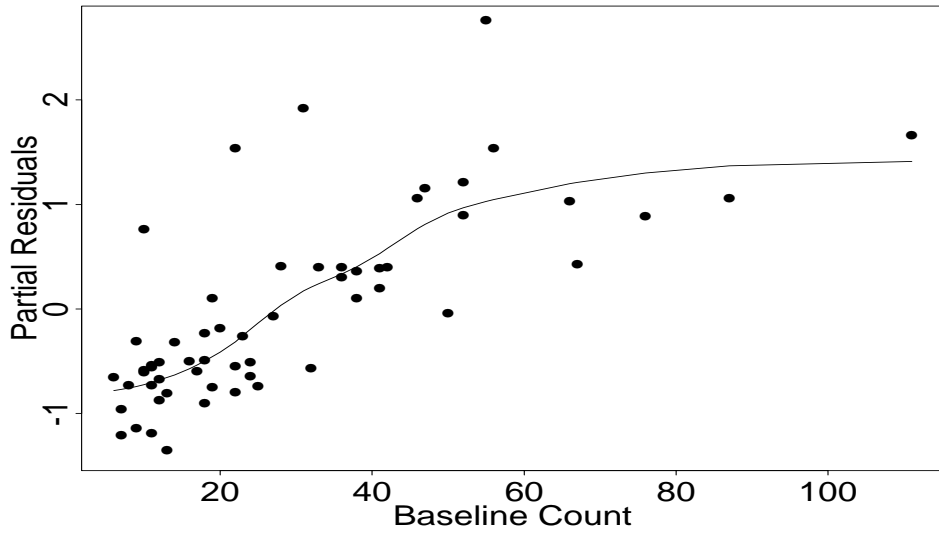


Figure 8: Partial residual plot for baseline count.

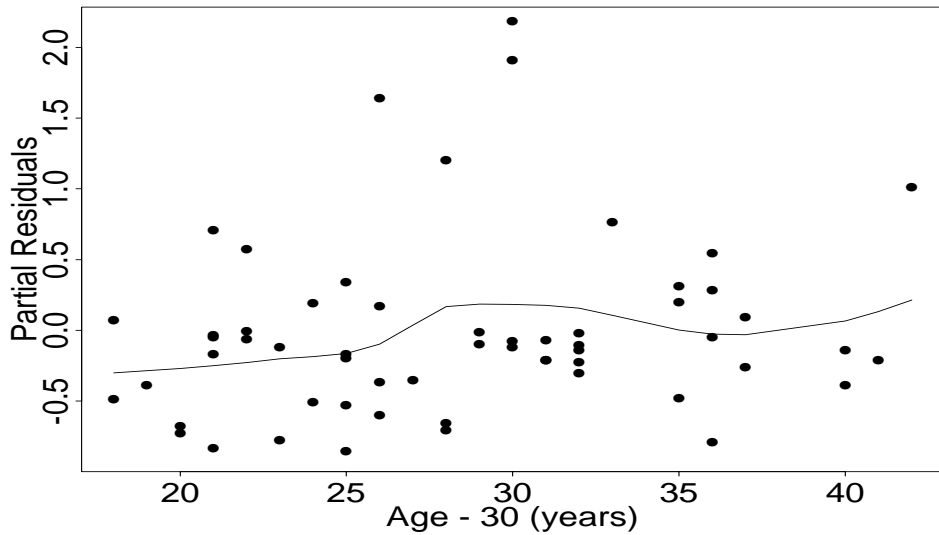


Figure 9: Partial residual plot for age.

between age and log baseline count. In order to understand this interaction more clearly, we fit a *generalized additive model* in which the linear predictor η was replaced by a *loess* smoothed function of age and baseline count [7]. The effect of the baseline count on subsequent seizure counts is indeed more pronounced for older patients than for younger ones (Figure 10).

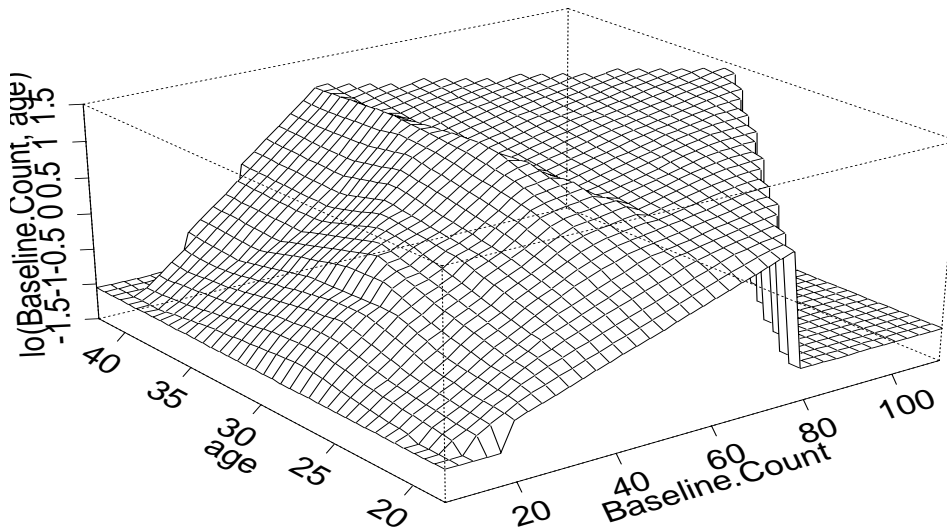


Figure 10: Smoothed (*loess*) response surface showing the interaction of baseline count and age.

5. OVERDISPERSION

A major assumption underlying the use of log-linear and logistic regression analysis for Poisson and binomially distributed data, respectively, is that the variance of the error distribution is completely determined by the mean. In practice this assumption often fails. The first indication that something is wrong is that the deviance measure of goodness-of-fit for “full” models exceeds its degrees of freedom. This phenomenon is known as *overdispersion*. We consider the following four aspects of the problem.

- Formal tests for overdispersion
- Standard errors for regression coefficients that account for overdispersion
- Test statistics for added variables that account for overdispersion
- More general models with parameters in the variance function

5.1 Testing for overdispersion

Table 1 shows the results of a “naive” Poisson regression analysis of the effect of Progabide on seizure rate that includes the two covariables and their interaction plus terms for a main and interactive treatment effect. The *t* statistics apparently indicate that the terms for Progabide and its interaction with the baseline count were highly statistically significant. The corresponding *score* and deviance tests for these treatment effects, both with 2 degrees of freedom, are 17.37 and 17.36. Any

conclusion that the treatment effects are real would be erroneous, however, since the residual deviance of 408.41 is much greater than the 52 degrees of freedom remaining after fitting a model with 6 parameters to 58 observations. The Poisson sampling model simply does not fit the data!

The overdispersion here is so great that there is really no doubt as to its presence. In other situations, however, one may wish a formal test that the extraneous variation is larger than predicted by the Poisson or binomial model. Score tests for variance parameters effectively compare the residuals with their expectation under the model [8, 9, 10]. For example, the adjusted score test of the hypothesis $H_0: \phi = 0$ for the *negative binomial* variance function $v(\mu; \phi) = \mu + \phi\mu^2$ is

$$T_1^2 = \frac{[\sum (y_i - \hat{\mu}_i)^2 - (1 - h_i)\hat{\mu}_i]^2}{2 \sum \hat{\mu}_i^2}$$

where the $\hat{\mu}_i$ denote the fitted values under H_0 , *i.e.*, under Poisson regression, and the h_i are the GLM *leverages*. The leverages account for the fact that the expectation of the residual sum of squares $\sum (y_i - \hat{\mu}_i)^2$ is slightly less than $\sum \mu_i$ due to estimation of the p regression coefficients. The score test of the hypothesis $H_0: \phi = 1$ in the model $v(\mu; \phi) = \phi\mu$ is

$$T_2^2 = \frac{1}{2n} \left\{ \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2 - (1 - h_i)\hat{\mu}_i}{\hat{\mu}_i} \right\}^2.$$

Both statistics have asymptotic χ_1^2 distributions under their respective null hypotheses. For the model in Table 1 they yield values of 1333.2 and 1411.0, respectively, confirming that the overdispersion is indeed immense. Dean [10] presents additional formulas for score tests for overdispersion in binomial regression models. Smith and Heitjan [11] developed related score tests for overdispersion in GLMs based on a model in which random effects are added to *each* of the p regression coefficients, not just the intercept. Their procedure is implemented in an S function available from Statlib at Carnegie Mellon University.

5.2 Adjusting standard errors for overdispersion

Even if the formal test is not statistically significant, the presence of overdispersion generally should be suspected whenever the deviance exceeds its expectation. Adjustment of the variance of the regression coefficients is essential in order to adequately reflect the uncertainty caused by the unexplained variation. The simplest adjustment is to multiply $\text{Var}(\hat{\beta})$ by an estimate of the GLM scale factor. The choices are $\hat{\phi} = D/(n - p)$ where D is the deviance or $\hat{\phi} = [\sum (y - \hat{\mu})^2 / \hat{v}] / (n - p)$ based on the Pearson statistic. Although strictly valid only if the true variances satisfy the GLM model assumption $\text{Var}(y) = \phi w^{-1} v(\mu)$, this method often yields a reasonable approximation more generally [11]. Alternatively, we may use the observed covariance matrix of the scores G instead of the expected information when calculating the asymptotic

variance [12], thus arriving at the *sandwich* variance estimate

$$\text{Var}_S(\hat{\beta}) = \mathcal{I}^{-1}G\mathcal{I}^{-1} = \mathcal{I}^{-1} \left[\sum_{i=1}^n w_i^2 \left(\frac{y_i - \mu_i}{v(\mu_i)} \right)^2 x_i x_i^t \right] \mathcal{I}^{-1}$$

where \mathcal{I} is the Fisher information *without* the scale factor ϕ (Appendix, equation 3) and μ_i is evaluated at the fitted value $\hat{\mu}_i$.

Table 1. Log-linear Poisson regression fit to the epilepsy data

Coefficient	Value	Std. Error	<i>t</i> -statistic
(Intercept)	3.079	0.451	6.833
log(Base.Cnt/4)	-0.073	0.201	-0.366
Age/10	-0.511	0.153	-3.332
log(Base.Cnt/4):Age/10	0.351	0.068	5.164
Progabide	-0.610	0.191	-3.197
Progabide:log(Base.Cnt/4)	0.204	0.088	2.325

Deviance=408.41; Pearson $\chi^2=456.52$; DF=52

With moderately sized samples, the sandwich variance often seems to underestimate the true variability. Thus it may be preferable to use a resampling variance estimate such as the *bootstrap* or its approximation the *jackknife* [13]. A close approximation to the latter is

$$\text{Var}_J = \frac{n-1}{n} \sum_{i=1}^n (\hat{\beta}_{(i)}^1 - \hat{\beta}_{(\cdot)}^1)(\hat{\beta}_{(i)}^1 - \hat{\beta}_{(\cdot)}^1)^t$$

where the $\hat{\beta}_{(i)}^1$ are the one-step deletion estimates defined in §3 and $\hat{\beta}_{(\cdot)}^1$ denotes their mean. Table 2 presents standard errors for the regression coefficients in Table 1 estimated by each of these methods. All suggest that there is considerably more uncertainty about the regression coefficients than was apparent from the “naive” standard errors shown in Table 1. Bootstrap standard errors are thought to be the most accurate [13]. The jackknife and sandwich standard errors underestimate, while the scaled Poisson standard errors overestimate, the variability with these data. Neither the main effect for Progabide, nor its interaction with the log baseline count, are statistically significant when appropriate account is taken of the overdispersion.

Table 2. Overdispersion adjusted standard errors for the Table 1 coefficients

Coefficient	Scale factor	Sandwich	1-step Jackknife	True Jackknife	Bootstrap (nb=5000)
(Intercept)	1.263	0.711	0.792	0.792	0.870
log(Base.Cnt/4)	0.564	0.326	0.368	0.369	0.424
Age/10	0.430	0.237	0.264	0.263	0.291
log(Base.Cnt/4):Age/10	0.190	0.104	0.117	0.117	0.137
Progabide	0.535	0.403	0.440	0.448	0.466
Progabide:log(Base.Cnt/4)	0.246	0.188	0.210	0.214	0.226

5.3 Adjusting test statistics for overdispersion

Tests of hypotheses of the form $H_0: \beta_2 = \beta_{20}$, where β_2 denotes a subset of the regression coefficients and β_{20} are a fixed set of values (usually zero), are useful for setting confidence bounds and for evaluating the importance of added variables in the regression model. The simplest is the Wald statistic $(\hat{\beta}_2 - \beta_{20})^t [\text{Var}_R(\hat{\beta}_2)]^{-1} (\hat{\beta}_2 - \beta_{20})$ where R is S or J . Using Var_S the two degree of freedom Wald test for a treatment effect with the epilepsy data is reduced from 17.23 to 2.89 ($p=0.24$). A version of the score test that replaces the usual Fisher information matrix with one based on the observed variability in the scores is also available [15]. Specifically, if $\hat{\beta}^0$ denotes the restricted estimate of the regression coefficients under H_0 and $U_2 = \partial \ell / \partial \beta_2$ is the vector of β_2 scores, the statistic is $U_2(\hat{\beta}^0)^t \Sigma^{-1} U_2(\hat{\beta}^0)$ where $\Sigma = G_{22} - \mathcal{I}_{21} \mathcal{I}_{11}^{-1} G_{12} - G_{21} \mathcal{I}_{11}^{-1} \mathcal{I}_{12} + \mathcal{I}_{21} \mathcal{I}_{11}^{-1} G_{11} \mathcal{I}_{11}^{-1} \mathcal{I}_{12}$ and the matrices \mathcal{I} and G have been partitioned conformally with β_2 . This yields the value 2.98 for the epilepsy data instead of the 17.37 obtained with the standard score test.

5.4 Parameters in the variance function

Table 3. Negative binomial regression fit to the epilepsy data

Coefficient	Value	Std. Error	t -statistic
(Intercept)	2.879	1.094	2.631
log(Base.Cnt/4)	0.003	0.570	0.005
Age/10	-0.399	0.368	-1.084
log(Base.Cnt/4):Age/10	0.303	0.191	1.588
Progabide	-0.702	0.451	-1.558
Progabide:log(Base.Cnt/4)	0.248	0.240	1.037

$\hat{\phi}=0.302$ by method of moments

If one has good *a priori* information that the variance function takes a particular parametric form, another approach to overdispersion is to fit the GLM corresponding to the specified variance. The method of moments [16] or pseudolikelihood [14] may be used to estimate the variance parameters. An attractive feature of the standard GLM model $\text{Var}(y) = \phi v(\mu)$ is that the value of ϕ has no effect on the estimate of β . Consequently, as illustrated above, all one need do is estimate ϕ after carrying out the standard log-linear Poisson or logistic regression analysis and use $\hat{\phi}$ to multiply the variances and divide the test statistics. This simple model is difficult to motivate on physical grounds, however. It is more natural to suppose that the extraneous variation, representing the effect of unmeasured covariables, arises from a random error term added to each of the linear predictors. For Poisson regression this leads to the negative binomial variance function $v(\mu; \phi) = \mu + \phi \mu^2$. Joint estimation of mean and variance parameters then involves iterating between the GLM score equations for β , assuming a particular value for ϕ , and moment or pseudolikelihood equations (see

Appendix) for ϕ . An S function for method-of-moments estimation and the *glm* procedure with a user-defined family corresponding to log link with the negative binomial variance were applied to the epilepsy data and led to the parameter estimates shown in Table 3. The results are broadly comparable with those obtained using Poisson regression with bootstrap or jackknife standard errors. Because of the uncertainty over the true variance function, however, the latter approach is usually preferable since its asymptotic validity requires only that the mean be correctly specified [6].

6. CORRELATION

One of the most critical GLM assumptions is that of statistical independence of the observations. This assumption is in doubt whenever there is a natural grouping or clustering of the data, in which case one needs to account for possible intra-cluster correlation when estimating the variance of the regression coefficients. The sandwich and resampling procedures just considered may be used for this purpose, except that the *units* used for calculating the empirical variances of the scores, for deletion in jackknife procedures or for sampling with replacement when performing bootstrap replications are the *clusters* rather than the individual observations. An explicit formula for the sandwich variance estimate for clustered data is given in the paper by Liang in this volume. We merely note here that it may be applied after a standard GLM fit, one that assumed independence of the data, by calling a simple S-function written by David Clayton. Likewise, S-functions for jackknife or bootstrap inference with independent observations [13] are easily modified to accommodate independent clusters of observations instead.

In order to illustrate these techniques we analyzed the epilepsy data once again, except that the model was fitted to 58 clusters of dependent response measurements consisting of the number of seizures during *each* of the four follow-up periods. Since the sum of independent Poisson observations is Poisson, the fit of the log-linear model to these 232 observations was identical to that shown in Table 1, with the exception that the intercept was estimated as 1.693, exactly $\log(4)$ less than earlier. Similarly, the sandwich and jackknife standard errors accounting for overdispersion and correlation as just described were identical to those shown in Table 2. A second model including an additional linear term, Visit, coded 1,2,3,4 for the four follow-up visits, yielded the estimates shown in Table 4. Accounting for overdispersion increased the estimated variance over the “naive” Poisson fit and accounting also for the intracluster correlation increased it even further. The only exception was for Visit, whose coefficient is determined in part by *within cluster* comparisons. The jackknife standard errors again exceeded those based on the sandwich.

Table 4. Model fit to overdispersed and correlated epilepsy data

Coefficient	Value	Standard Errors				
		Naive	Overdispersion		Correlation	
			Sandwich	Jackknife ¹	Sandwich	Jackknife ²
(Intercept)	1.799	0.454	0.583	0.606	0.730	0.814
log(Base.Cnt/4)	-0.074	0.201	0.277	0.292	0.325	0.369
Age/10	-0.511	0.153	0.182	0.190	0.237	0.263
log(Base.Cnt/4):Age/10	0.351	0.068	0.089	0.094	0.104	0.117
Progabide	-0.610	0.191	0.293	0.301	0.402	0.448
Progabide:log(Base.Cnt/4)	0.204	0.088	0.140	0.144	0.187	0.214
Visit/10	-0.427	0.220	0.388	0.406	0.377	0.395

¹ One-step jackknife; ² True jackknife

7. CHECKING THE LINK FUNCTION

The final question concerns the appropriateness of the link function g . In Poisson regression, for example, we may wish to examine whether the usual log link $\eta = \log(\mu)$ is appropriate or whether the simply identity $\eta = \mu$, under which effects of covariables combine additively rather than multiplicatively, might serve instead. For binomial regression the choice may be between the logit link and the complementary log-log link $\eta = \log[-\log(1-\mu)]$ which allows for asymmetry about $\mu = 1/2$.

Pregibon [17] advocated the comparison of two link functions by imbedding them both in a parametric family of link functions $g(\mu; \lambda)$. The Box-Cox family of power transforms $g(\mu; \lambda) = (\mu^\lambda - 1)/\lambda$ yields the log link at $\lambda = 0$ and the identity at $\lambda = 1$. Likewise, the family $g(\mu; \lambda) = \log\{[(1-\mu)^{-\lambda} - 1]/\lambda\}$ gives the logit link at $\lambda=1$ and the c-log-log link at $\lambda=0$. Pregibon's goodness-of-link test consists of a score test of the hypothesis $H_0: \lambda = \lambda_0$ which is easily implemented using the method of *constructed variables*. After fitting the model with the standard link ($\lambda = \lambda_0$), and obtaining the fitted values $\hat{\mu}$, one tests the significance of the constructed variable

$$z = - \left. \frac{\partial g(\hat{\mu}; \lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0}$$

using the standard score test for added variables.

For testing the log link imbedded in the Box-Cox power family, two applications of l'Hôpital's rule give $z = -\log^2(\hat{\mu})/2$. We applied this goodness-of-link test to the model shown in Table 1, using the robust version of the score test, to find $\chi_1^2 = 0.30$ ($p = 0.58$). The analogous test based on the negative binomial model (Table 3) yielded the value 0.80 ($p = 0.37$). Thus there is no evidence to suggest a problem with the log link for this problem.

The regression coefficient of the constructed variable z provides the change in λ when using Fisher scoring to find the maximum likelihood estimate of λ . Further uses

of constructed variables to handle other non-linear parameters, for example in the covariables, are described in Chapter 11 of McCullagh and Nelder [1].

8. SUMMARY AND CONCLUSIONS

Fitting of GLMs and other regression models involve a number of assumptions that often are not adequately appreciated or evaluated by the data analyst. Graphical and analytical procedures are available that allow one to check many of the model assumptions and to make adjustments when they fail to hold. Residual plots and one-step deletion diagnostics are useful for detecting outliers and “high leverage” observations, respectively. Partial residual plots and univariate or bivariate regression smooths [2, 7] may suggest appropriate transformations of covariables and help to visualize the nature of interactions. Even if score tests for overdispersion do not show statistical significance, its presence should be suspected. The one-step jackknife variance estimate is just as easy to obtain with modern software as the popular sandwich estimate. Both it and the bootstrap often yield more realistic measures of uncertainty in small samples. Bootstrapping or jackknifing a GLM fit using the cluster as the sampling unit accounts for correlation as well as overdispersion. The method of constructed variables is useful for estimation and testing of non-linear parameters such as those in the link function. These methods greatly extend the scope of GLMs and help to avoid unwarranted inferences being drawn from uncritical use of the standard procedures.

ACKNOWLEDGEMENTS

This work was supported in part by grant 5 R01 CA40644 from the US National Institutes of Health, Bethesda, MD.

References

- [1] McCullagh P and Nelder JA (1989) *Generalized Linear Models. Second Edition* London: Chapman and Hall.
- [2] Chambers JM and Hastie TJ (1992) *Statistical Models in S* Pacific Grove, California: Wadsworth & Brooks.
- [3] Wedderburn RWM (1974) Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* 61:439-447.
- [4] Thall PF and Vail SC (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics* 46:657-671.
- [5] Cook RP and Weisberg S (1982) *Residuals and Influence in Regression*. London: Chapman and Hall.

- [6] Diggle PJ, Liang K-Y and Zeger SL (1994) *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- [7] Hastie T and Tibshirani R (1990). *Generalized Additive Models*. London: Chapman and Hall.
- [8] Breslow N (1989) Score tests in overdispersed GLMs. *Workshop on Statistical Modelling* (Decarli A, Francis BJ, Gilchrist R, Seeber GUH, eds) New York: Springer, pp. 64-74.
- [9] Dean C, Lawless JF (1989) Tests for detecting overdispersion in Poisson regression models. *J Am Statist Assoc* 84:467-472.
- [10] Dean C (1992) Testing for overdispersion in Poisson and binomial regression models. *J Am Statist Assoc* 87: 451-457.
- [11] Smith PJ, Heitjan DF (1993) Testing and adjusting for departures from nominal dispersion in Generalized Linear Models. *Appl. Statist.* 42:31-41.
- [12] Huber PJ (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 221-233. Berkeley: University of California Press.
- [13] Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall.
- [14] Davidian M, Carroll RJ (1987) Variance function estimation. *J. Am. Statist. Assoc.* 82:1079-1091.
- [15] Breslow N (1990) Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Am. Statist. Assoc.* 85:665-571.
- [16] Breslow N (1984) Extra Poisson variation in log-linear models. *Appl. Statist.* 33:38-44.
- [17] Prebignon D (1980) Goodness-of-link tests for generalized linear models. *Appl. Statist.* 29:15-24.

APPENDIX: A REVIEW OF GLM THEORY

The standard theoretical approach to the GLM is via maximum likelihood estimation in exponential dispersion models. The log-likelihood function is

$$\ell(\beta) = \frac{1}{\phi} \sum_{i=1}^n w_i [y_i \theta_i(\beta) - b(\theta_i(\beta))] + \sum_{i=1}^n c(y_i, \phi, w_i) \quad (1)$$

where b is the cumulant generating function for the particular exponential family distribution and θ , the *canonical parameter* in the GLM, is defined by the relation $\mu = b'(\theta)$, *i.e.*, as the inverse function of $b'(\cdot)$ evaluated at μ . The identity $v(\mu) = b''(\theta(\mu))$ shows that the particular exponential family distribution may be specified either by b or by v . Thus the Poisson distribution has $b(\theta) = \exp(\theta)$ and $v(\mu) = \mu$ while the binomial has $b(\theta) = \log(1 + e^\theta)$ and $v(\mu) = \mu(1 - \mu)$ [1]. The theory of quasi-likelihood extends GLM techniques to variance functions that do not correspond to an exponential family distribution [3].

The *deviance* measure d_i of the discrepancy between the observation y_i and its fitted mean $\hat{\mu}_i$ is defined by

$$d_i(y; \hat{\mu}) = 2\phi[\ell_i(y_i; y_i) - \ell_i(\hat{\mu}_i; y_i)] = -2w_i \int_{y_i}^{\hat{\mu}_i} \frac{y_i - u}{v(u)} du,$$

where $\ell_i(\mu; y)$ is the log-likelihood contribution for the i^{th} observation, considered as a function of the mean. Minimization of the overall deviance $D = \sum_i d_i(y_i; \hat{\mu}_i)$ is equivalent to maximization of the log-likelihood (1). *Deviance residuals* $r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$ behave roughly like normal theory residuals when the model holds and are widely used to identify outliers and for other model checking purposes.

Estimates of the regression coefficients are found by equating to zero the *score* vector

$$U(\beta) = \frac{\partial \ell}{\partial \beta} = \frac{1}{\phi} \sum_{i=1}^n w_i (y_i - \mu_i) \frac{\partial \theta_i}{\partial \beta} = \frac{1}{\phi} \sum_{i=1}^n w_i \left(\frac{y_i - \mu_i}{v_i} \right) \frac{\partial \mu_i}{\partial \beta} = \frac{1}{\phi} \sum_{i=1}^n w_i \left(\frac{y_i - \mu_i}{v_i g'(\mu_i)} \right) x_i, \quad (2)$$

where $v_i = v(\mu_i)$. For the *canonical link* functions $g(\mu) = \theta(\mu)$, these being the usual log and logit transforms for Poisson and binomial data, $g'(\mu) = 1/v(\mu)$ and the score equations are the standard normal equations $\sum_i w_i (y_i - \mu_i) x_i = 0$. The usual first order asymptotic properties of the resulting estimates $\hat{\beta}$, and of test statistics based on the scores and deviance differences, depend only on correct specification of the variance function and not on the higher moments implied by the corresponding exponential family distribution.

The asymptotic variance of $\hat{\beta}$ is obtained as the inverse of the Fisher information matrix

$$\mathcal{I}(\beta) = -\mathbb{E} \frac{\partial^2 \ell}{\partial \beta \partial \beta^t} = \frac{1}{\phi} \sum_{i=1}^n w_i v_i^{-1} \frac{\partial \mu_i}{\partial \beta} \frac{\partial \mu_i}{\partial \beta^t} = \frac{1}{\phi} \sum_{i=1}^n w_i v_i \frac{\partial \theta_i}{\partial \beta} \frac{\partial \theta_i}{\partial \beta^t} = \frac{1}{\phi} \sum_{i=1}^n \frac{w_i}{v_i [g'(\mu_i)]^2} x_i x_i^t, \quad (3)$$

evaluated at the fitted values $\hat{\mu}_i$. The key GLM computational feature is that use of Fisher scoring to find $\hat{\beta}$ is equivalent to repeated weighted least squares regression of the *working vector* with components $Y_i = x_i^t \beta + (y_i - \mu_i) g'(\mu_i)$ on the explanatory variables x_i using as weights the GLM *iterated weights* $W_i = w_i \{v_i [g'(\mu_i)]^2\}^{-1}$. $W_i = w_i v_i$ for models with canonical link. This permits use of standard least squares techniques

for model checking to be extended to the GLM. For example, the one-step approximations of §4 are given by $\hat{\beta}_{(i)}^1 = (X^t W X)^{-1} W_i x_i r_i^W / (1 - h_i)$ where W is diagonal with elements W_i , r_i^W denotes the *working residual* $r_i^W = Y_i - x_i \hat{\beta} = (y_i - \hat{\mu}_i) g'(\hat{\mu}_i)$ and the *leverages* h_i are the diagonal elements of the *hat* matrix $H = X(X^t W X)^{-1} X^t W$. The leverages h_i and one-step estimates $\hat{\beta}_{(i)}^1$ are obtained by applying the *lm.influence* function to the results of a *glm* fit in S. Quasilikelihood score tests for added variables are obtained by applying the *add1* function.

Suppose there are no prior weights and that the true variance is $\text{Var}(y) = v = v(\mu, \phi)$ where ϕ is a variance parameter or, possibly, a vector of variance parameters. Using quasilikelihood to estimate β and psuedolikelihood [14] to estimate ϕ we jointly solve the two sets of estimating equations

$$U(\beta, \phi) = \sum u = \sum \frac{y - \mu}{v} \cdot \frac{\partial \mu}{\partial \beta} = 0$$

and

$$\tilde{U}(\beta, \phi) = \sum \tilde{u} = \sum \frac{(y - \mu)^2 - (1 - h)v}{v^2} \cdot \frac{\partial v}{\partial \phi} = 0.$$

Tests for overdispersion based on the second equation lead to the test statistics of §5 [8].