Pearson's Goodness of Fit Statistic as a Score Test Statistic*

Gordon K. Smyth

Abstract

For any generalized linear model, the Pearson goodness of fit statistic is the score test statistic for testing the current model against the saturated model. The relationship between the Pearson statistic and the residual deviance is therefore the relationship between the score test and the likelihood ratio test statistics, and this clarifies the role of the Pearson statistic in generalized linear models. The result is extended to cases in which there are multiple reponse observations for the same combination of explanatory variables.

Keywords. Pearson statistic; score test; chisquare statistic; generalized linear model; exponential family nonlinear model; saturated model.

1 Introduction

Goodness of fit tests go back at least to Pearson's (1900) article establishing the asymptotic chisquare distribution for a goodness of fit statistic for the multinomial distribution. Pearson's chisquare statistic includes the test for independence in two-way contingency tables. It has been extended in generalized linear model theory to a test for the adequacy of the current fitted model. Given a generalized linear model with responses y_i, weights w_i, fitted means $\hat{\mu}_i$, variance function $v(\mu)$ and dispersion $\phi = 1$, the Pearson goodness of fit statistic is

$$X^2 = \sum \frac{w_i(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}$$

[14]. If the fitted model is correct and the observations y_i are approximately normal, then X^2 is approximately distributed as χ^2 on the residual degrees of freedom for the model.

*To appear in *Science and Statistics: A Festschrift for Terry Speed*, D. R. Goldstein (ed.), IMS Lecture Notes — Monograph Series, Volume 40, Institute of Mathematical Statistics, Hayward, California, March 2003.

The Pearson goodness of fit statistic X^2 is one of two goodness of fit tests in routine use in generalized linear models, the other being the residual deviance. The residual deviance is the log-likelihood ratio statistic for testing the fitted model against the saturated model in which there is a regression coefficient for every observation. The Pearson statistic is a quadratic form alternative to the residual deviance, and is often preferred over the residual deviance because of its moment estimator character. The expected value of the Pearson statistic depends only on the first two moments of the distribution of the y_i and in this sense the Pearson statistic is robust against mis-specification of the response distribution.

The score test, like the likelihood ratio test, is a general asymptotic parametric test associated with the likelihood function [22]. Score tests are often simpler than likelihood ratio tests because the statistic requires parameter estimators to be obtained only under the null hypothesis. For this reason score tests have been proposed frequently in generalized linear model contexts to test for various sorts of model complications such as overdispersion [5] [3] [7] [24] [13] [19], zero inflation [8], adequacy of the link function [20] [9], or extra terms in the fitted model [21] [4] [1] [2] [26] [19].

While the residual deviance arises from a general inferential principle, namely the likelihood ratio test, the origin of the Pearson statistic has seemed more ad hoc. Several authors have noted that score tests for extra terms in the linear predictor give rise to chisquare statistics, but there has been no result for the residual Pearson statistic itself. Pregibon [21] shows, by using one-step estimators, that the score statistic for extra terms in the linear predictor can be expressed as a difference between two chisquare statistics, just as the likelihood ratio test can be obtained as the difference between two residual deviances. Cox and Hinkley [6, Examples 9.17 and 9.21] show that the simplest Pearson statistic, the goodness of fit statistic for the multinomial distribution, can be derived as a score statistic. This article shows that Cox and Hinkley's result for the multinomial extends to all generalized linear models. The Pearson goodness of fit statistic is itself a score test statistic, testing the current model against the saturated model. The relationship between the Pearson statistic and the residual deviance is therefore the relationship between the score test and the likelihood ratio test statistics, and this clarifies the role of the Pearson statistic in generalized linear models.

The result of this article extends to several more general situations. The result extends to data sets with multiple counts in categories and to generalizations of exponential families models, such as overdispersion models, for which there are extra parameters in the variance function. It includes for example as special cases the results on tests for independence in two-way contingency tables of Thall [26] and Paul and Banerjee [19]. The general proofs given here are simpler and more transparent than the special case proofs for contingency tables. Finally, the results given here do not require link-linearity as in generalized linear models, but apply to any exponential family non-linear regression model.

The theory of score tests is revised briefly in Section 2 and the background material required for generalized linear and non-linear models is stated briefly in Section 3. The main results of the article are given in Section 4 showing the relationship between score tests and goodness of fit. Section 5 goes on to consider models with extra-dispersion and Section 6 considers estimation of the dispersion parameter.

2 Score tests

This section summarizes briefly the theory of likelihood score tests. Further background on score tests and likelihood ratio tests can be found in Rao [23, pages 417–418] and Cox and Hinkley [6, Section 9.3]. Let $\ell(\mathbf{y};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ be a log-likelihood function depending on a response vector $\mathbf{y}$ and parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. We wish to test the composite hypothesis $H_0 : \boldsymbol{\theta}_2 = 0$ against the alternative that $\boldsymbol{\theta}_2$ is unrestricted. The components of $\boldsymbol{\theta}_1$ are so-called nuisance parameters because they are not of interest in the test but values must be estimated for them for a test statistic to be computed. The likelihood score vectors for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the partial derivatives

$$\dot{\ell}_1 = \frac{\partial \ell}{\partial \boldsymbol{\theta}_1}$$

and

$$\dot{\ell}_2 = \frac{\partial \ell}{\partial \boldsymbol{\theta}_2}$$

respectively. The observed information matrix for the parameters is $-\ddot{\ell}$ with

$$\ddot{\ell} = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_2^T} = \left(\begin{array}{cc} \ddot{\ell}_{11} & \ddot{\ell}_{12} \\ \ddot{\ell}_{21} & \ddot{\ell}_{22} \end{array} \right).$$

The expected or Fisher information matrix is $\mathcal{I} = E(-\ddot{\ell})$, which is partitioned conformally with $\ddot{\ell}$ as

$$\mathcal{I} = \left(\begin{array}{cc} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{array} \right).$$

The score test statistic is based on the fact that the score vector $\dot{\ell}$ has mean zero and covariance matrix $\mathcal{I}$. If the nuisance vector $\boldsymbol{\theta}_1$ is known, then the score test statistic of H_0 is

$$Z = \mathcal{I}_{22}^{-1/2} \dot{\ell}_2,$$

where $\mathcal{I}_{22}^{1/2}$ stands for any factor such that $\mathcal{I}_{22}^{1/2}\mathcal{I}_{22}^{T/2} = \mathcal{I}_{22}$, or equivalently

$$S = Z^T Z = \dot{\ell}_2^T \mathcal{I}_{22}^{-1} \dot{\ell}_2$$

with ℓ_2 and $\mathcal{I}_{22}$ evaluated at $\boldsymbol{\theta}_2 = 0$. The score vector $\dot{\ell}$ is a sum of terms corresponding to individual observations and so is asymptotically normal under standard regularity conditions. It follows that Z is asymptotically a standard normal p_2-vector under the null hypothesis H_0 and that S is asymptotically chisquare distributed on p_2 degrees of freedom, where p_2 is the dimension of $\boldsymbol{\theta}_2$.

If the nuisance parameters are not known, then the score test substitutes for them their maximum likelihood estimators $\hat{\boldsymbol{\theta}}_1$ under the null hypothesis. Setting $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1$ is equivalent to setting $\dot{\ell}_1 = 0$, so we need the asymptotic distribution of $\dot{\ell}_2$ conditional on $\dot{\ell}_1 = 0$, which is normal with mean zero and covariance matrix

$$\mathcal{I}_{2.1} = \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}.$$

The score test statistic becomes

$$S = \dot{\ell}_2^T \mathcal{I}_{2.1}^{-1} \dot{\ell}_2$$

with $\dot{\ell}_2$ and $\mathcal{I}_{2.1}$ evaluated at $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_2 = 0$. If $\mathcal{I}_{12} = 0$ then $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be orthogonal. In that case, $\dot{\ell}_1$ and $\dot{\ell}_2$ are independent and $\mathcal{I}_{2.1} = \mathcal{I}_{22}$, meaning that the information matrix $\mathcal{I}_{22}$ does not need to be adjusted for estimation of $\boldsymbol{\theta}_1$,

Neyman [15] and Neyman and Scott [16] show that the asymptotic distribution and efficiency of the score statistic S is unchanged if an estimator other than the maximum likelihood estimator is used for the nuisance parameters, provided that the estimator is consistent with convergence rate at least $O(n^{-1/2})$, where n is the number of observations. They show that we can substitute into S any estimator $\tilde{\boldsymbol{\theta}}_1$ of $\boldsymbol{\theta}_1$ for which $\sqrt{n}|\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1|$ is bounded in probability as $n \to \infty$. In that case they rename the score statistic the $C(\alpha)$ test statistic.

3 Generalized Linear Models

Generalized linear models assume that observations are distributed according to a linear exponential family with an additional dispersion parameter. The density or probability mass function for each response is assumed to be of the form

$$f(y; \mu, \phi) = a(y, \phi) \exp[\{y\theta - \kappa(\theta)\}/\phi] \tag{1}$$

where a and κ are suitable known functions. The mean is $\mu = \dot{\kappa}(\theta)$ and the variance is $\phi\ddot{\kappa}(\theta)$. The mean μ and the canonical parameter θ are one-to-one functions of one another. We call ϕ the dispersion parameter and $v(\mu) = \ddot{\kappa}(\theta)$ the variance function.

Following Jørgensen [10, 12], we call the distribution described by (1) an exponential dispersion model and denote it $\mathrm{ED}(\mu, \phi)$. If $y_1, \ldots, y_n$ are independently distributed as $\mathrm{ED}(\mu, \phi)$ then the sample mean $\bar{y}$ is sufficient for μ and $\bar{y} \sim \mathrm{ED}(\mu, \phi/n)$. More generally, if $y_i \sim \mathrm{ED}(\mu, \phi/w_i)$ where the w_i are known weights, then the weighted sum $\bar{y}_w$ is sufficient for μ and

$$\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \sim \mathrm{ED}\left(\mu, \frac{\phi}{\sum_{i=1}^n w_i}\right).$$

A generalized linear model assumes independent observations $y_1, \dots, y_n$ with $y_i \sim \mathrm{ED}(\mu_i, \phi/w_i)$. The means μ_i are assumed to follow a link-linear model

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{2}$$

where g is a known monotonic link function, $\mathbf{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients. Without loss of generality we will assume that the $n \times p$ matrix X with rows $\mathbf{x}_i$ is of full column rank and that $p < n$, where p is the dimension of $\boldsymbol{\beta}$.

More generally we will consider generalized nonlinear models in which the mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ is a general n-dimensional function of the p-vector $\boldsymbol{\beta}$. To ensure that the parametrization is not degenerate, we will assume that the gradient matrix $\partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ is of full column rank, at least in a neighborhood containing the true value of $\boldsymbol{\beta}$ and the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

In this article we mainly consider models in which the dispersion is known, $\phi = 1$ say. Most models with discrete responses have known dispersion.

4 Goodness of Fit Tests

Let Ω be the locus of possible values for $\boldsymbol{\mu}$, $\Omega = \{\boldsymbol{\mu}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Let H_0 be the null hypothesis that $\boldsymbol{\mu}$ belongs to Ω and let H_a be the alternative hypothesis that $\boldsymbol{\mu}$ is unrestricted. The goodness of fit test for the current model tests H_0 against H_a. For a generalized linear model, H_0 is the hypothesis that the μ_i are described by the link-linear model (2).

Theorem 1
The score statistic for the goodness of fit test of a generalized nonlinear model with unit dispersion is the Pearson chisquare statistic

$$S = \sum_{i=1}^{n} w_i (y_i - \hat{\mu}_i)^2 / v(\hat{\mu}_i)$$

where $\hat{\mu}_i$ is the expected value μ_i evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$.

Proof. There exists an parameter vector $\boldsymbol{\beta}_2$ of dimension $n-p$ such that $(\boldsymbol{\beta}, \boldsymbol{\beta}_2)$ is a one-to-one transformation of $\boldsymbol{\mu}$ in the neighborhood of interest and such that $\boldsymbol{\beta}_2 = 0$ if and only if $\boldsymbol{\mu} \in \Omega$. The goodness of fit test consists of testing $H_0 : \boldsymbol{\beta}_2 = 0$ against the alternative that $\boldsymbol{\beta}_2$ is unrestricted. The components of the original parameter vector $\boldsymbol{\beta}$ are the nuisance parameters for this test. In the generalized linear model case, the implicit parameter vector $\boldsymbol{\beta}_2$ can be constructed by finding an $n \times (n-p)$ matrix X_2 such that (X, X_2) is of full rank. Then H_a is the saturated model that $g(\mu_i) = X\boldsymbol{\beta} + X_2\boldsymbol{\beta}_2$ for some $\boldsymbol{\beta}$ and some $\boldsymbol{\beta}_2$.

Let $\dot{\ell}_1$ and $\dot{\ell}_2$ be the score vectors for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_2$ respectively, and let $\mathcal{I}$ be the Fisher information matrix, partitioned into $\mathcal{I}_{11}$, $\mathcal{I}_{12}$ and $\mathcal{I}_{22}$ as in Section 2. The Fisher information for $\boldsymbol{\beta}_2$ adjusted for estimation of $\boldsymbol{\beta}$ is $\mathcal{I}_{2.1}$ and the score statistic for testing H_0 is

$$S = \dot{\ell}_2^T \mathcal{I}_{2.1}^{-1} \dot{\ell}_2$$

where $\dot{\ell}_2$ and $\mathcal{I}_{2.1}$ are evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_2 = 0$.

Let $V = \mathrm{diag}\{v(\mu_i)/w_i\}$ and write

$$\mathbf{e} = V^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$$

for the vector of Pearson residuals. Also write

$$U_1 = V^{-1/2} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$$

and

$$U_2 = V^{-1/2} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}_2}.$$

It is straightforward to show that the score vectors are given by

$$\dot{\ell}_j = U_j^T \mathbf{e}$$

for $j = 1, 2$ and the information matrices are given by

$$\mathcal{I}_{jk} = U_j^T U_k$$

for $j, k = 1, 2$ [25] [27].

Write P_1 for the matrix $P_1 = U_1(U_1^T U_1)^{-1} U_1^T$ of the orthogonal projection onto the column space of U_1. Also write

$$U_{2.1} = (I - P_1)U_2$$

and $P_{2.1}$ for the matrix

$$P_{2.1} = U_{2.1}(U_{2.1}^T U_{2.1})^{-1} U_{2.1}^T$$

of the orthogonal projection onto the column space of $U_{2.1}$. Then P_1 and $P_{2.1}$ project onto orthogonal subspaces and $P_1 + P_{2.1} = I$ since the dimensions of the subspaces add to n.

We can rewrite

$$\mathcal{I}_{2.1} = U_2^T U_2 - U_2^T U_1 (U_1^T U_1)^{-1} U_1^T U_2 = U_2^T (I - P_1) U_2 = U_{2.1}^T U_{2.1}.$$

We can also rewrite

$$\dot{\ell}_2 = (U_2^T - U_2^T P_1)\mathbf{e} = U_{2.1}^T \mathbf{e}$$

because evaluating at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ ensures that $U_1^T \mathbf{e} = 0$ and hence $P_1 \mathbf{e} = 0$. This shows that the score statistic is

$$S = \mathbf{e}^T U_{2.1} (U_{2.1}^T U_{2.1})^{-1} U_{2.1}^T \mathbf{e} = \mathbf{e}^T P_{2.1} \mathbf{e} = \mathbf{e}^T (P_1 + P_{2.1}) \mathbf{e} = \mathbf{e}^T \mathbf{e}$$

which the Pearson statistic. □

Example. Theorem 1 shows that the chisquare test for independence in a twoway contingency table is a score statistic, based on the assumption that the counts are independent and Poisson distributed. For multiway contingency tables, Theorem 1 shows that the score test of the hypothesis that any chosen subset of the pairs of faces are independent yields a Pearson statistic.

We now consider the case where there are multiple observations for some or all of the covariate combinations. In such cases it is usually more natural to associate the saturated alternative with unique combinations of the explanatory variables rather than to allow every μ_i to be different. The following corollary to Theorem 1 shows that the score test statistic in such cases is naturally expressed in terms of the mean response for each covariate-combination group. The score statistic in the corollary is the Pearson goodness of fit statistic when the data has been reduced to sufficient statistics for each covariate-combination group.

Corollary 1
Suppose that $y_{ij} \sim \mathrm{ED}(\mu_i, 1/w_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, *are independent. The score test statistic of* H_0, *that the* μ_i *are functions of* $\boldsymbol{\beta}$, *against the alternative* H_a *that they are unrestricted, is given by*

$$S = \sum_{i=1}^{n} w_{i\cdot}(\bar{y}_{wi} - \hat{\mu}_i)^2 / v(\hat{\mu}_i)$$

where $\hat{\mu}_i$ *is the maximum likelihood estimator of* μ_i, $w_{i\cdot}$ *is the sum of weights*

$$w_{i\cdot} = \sum_{j=1}^{n_i} w_{ij}$$

and $\bar{y}_{wi}$ *is the weighted mean*

$$\bar{y}_{wi} = \frac{1}{w_{i\cdot}} \sum_{j=1}^{n_i} w_{ij} y_{ij}.$$

Proof. The weighted means $\bar{y}_{wi}$ are sufficient for the μ_i, and $\bar{y}_i \sim \mathrm{ED}(\mu_i, 1/w_{i\cdot})$. The $\bar{y}_{wi}$ are distributed as for the y_i but with weights $w_{i\cdot}$, so the result follows immediately from Theorem 1. □

Example. Suppose that the y_{ij} are binary responses and that $w_{ij} = 1$ for all i and j. Then

$$S = \sum_{i=1}^{n} n_i (r_i - \hat{p}_i)^2 / v(\hat{p}_i)$$

where r_i is the empirical proportion for the ith covariate-combination group, $\hat{p}_i$ is the estimated probability that $y_{ij} = 1$, and $v(\hat{p}_i) = \hat{p}_i(1 - \hat{p}_i)$. If $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$ is the number of successes for the ith group then the $y_{i\cdot}$ are binomial(n_i, p_i) and

$$S = \sum_{i=1}^{n} (y_{i\cdot} - \hat{\mu}_i)^2 / v_i(\hat{\mu}_i)$$

with $\mu_i = np_i$ and $v_i(\mu_i) = \mu_i(n_i - \mu_i)/n_i$. This is the Pearson goodness of fit statistic for the data summarized in the usual generalized linear model way as binomial counts for each covariate-combination group.

Example. Paul and Banerjee [19] derive the score test for interaction in a twoway contingency table with multiple counts in each cell. Corollary 1 includes Paul and Banerjee's Theorem 1 as a special case.

5 Extra Parameters in the Variance

Suppose now that there are extra parameters which affect the variance of the y_i, but not its mean, and which are outside the exponential dispersion model framework. Let $\boldsymbol{\gamma}$ be the vector of extra parameters and let G be the parameter space for $\boldsymbol{\gamma}$. Suppose that for each fixed value of $\boldsymbol{\gamma}$, the y_i follow a generalized nonlinear model with variance function $\mu \to v(\mu; \boldsymbol{\gamma})$. The values of $\boldsymbol{\gamma}$ effectively index a class of generalized nonlinear models. This setup arises frequently when extra parameters are introduced to accommodate overdispersion in generalized linear models [1] [2] [7] [19].

It is straightforward to show that $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are orthogonal parameters. This follows because

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} V^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

and $\boldsymbol{\mu}$ does not depend on $\boldsymbol{\gamma}$. Therefore the cross derivative $\partial^2 \ell / \partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}$ will be linear in $\mathbf{y} - \boldsymbol{\mu}$ and will have expectation zero.

Orthogonality of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ implies that estimation of $\boldsymbol{\gamma}$ does not affect the form of the score statistics for goodness of fit. According to the theory of $C(\alpha)$ tests, $\boldsymbol{\gamma}$ may be replaced in the score test statistics by any estimator which is $O(n^{-1/2})$ consistent without changing the distributional properties of S to first order. This gives the following theorem.

Theorem 2
Suppose that for each $\boldsymbol{\gamma} \in G$, $y_1, \ldots, y_n \sim \mathrm{ED}(\mu_i, 1/w_i)$ are independent with variance function $v(\mu; \boldsymbol{\gamma})$. The $C(\alpha)$ goodness of fit statistic is

$$S = \sum_{i=1}^{n} w_i (y_i - \hat{\mu}_i)^2 / v(\hat{\mu}_i; \tilde{\boldsymbol{\gamma}})$$

where $\tilde{\boldsymbol{\gamma}}$ is any $\sqrt{n}$-consistent estimator of $\boldsymbol{\gamma}$ and $\hat{\mu}_i$ is the maximum likelihood estimator of μ_i given $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$.

Corollary 2
Suppose that for each $\boldsymbol{\gamma} \in G$, $y_{ij} \sim \mathrm{ED}(\mu_i, 1/w_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, are independent with variance function $v(\mu; \boldsymbol{\gamma})$. The $C(\alpha)$ goodness of fit statistic is

$$S = \sum_{i=1}^{n} w_{i\cdot}(\bar{y}_{wi} - \hat{\mu}_i)^2 / v(\hat{\mu}_i; \tilde{\boldsymbol{\gamma}})$$

where $\tilde{\boldsymbol{\gamma}}$ is any $\sqrt{n}$-consistent estimator of $\boldsymbol{\gamma}$, $\hat{\mu}_i$ is the maximum likelihood estimator of μ_i given $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$, the $w_{i\cdot}$ are sums of weights and the $\bar{y}_{wi}$ are weighted means.

The proofs of Theorem 2 and the corollary are similar to the proofs in Section 2.

Example. Suppose that y_{ij} follows a negative binomial distribution with mean μ_i and variance function $V(\mu; c) = \mu + c\mu^2$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ for each $c \geq 0$. Suppose that the μ_i are a function of a vector $\boldsymbol{\beta}$ of regression parameters. For fixed values of c, the means $\bar{y}_i$ are sufficient for the μ_i and are negative binomial with the same variance function and weights n_i. The $C(\alpha)$ goodness of fit statistic therefore is

$$S = \sum_{i=1}^{n} \frac{n_i(\bar{y}_i - \hat{\mu}_i)^2}{\hat{\mu}_i + \tilde{c}\hat{\mu}_i^2}$$

where $\tilde{c}$ is a $\sqrt{n}$-consistent estimator of c and $\hat{\mu}_i$ is the maximum likelihood estimator of μ_i with $c = \tilde{c}$. This includes Theorem 3 of Paul and Banerjee (1998).

One possible estimator for $\boldsymbol{\gamma}$ is the maximum likelihood estimator. An alternative estimation method is to solve $S = n - p$ with respect to $\boldsymbol{\gamma}$. This estimator is often preferred in overdispersion contexts because it is evidently a consistent estimator based only on the first and second moments of the y_i and therefore has a quasi-likelihood flavor (Breslow, 1990). Obviously the score statistic S is no longer useful as a goodness of fit statistic if $\boldsymbol{\gamma}$ is estimated by either of the above methods.

If there are repeat observations for covariate combinations, then an estimate of $\boldsymbol{\gamma}$ may be obtained from the 'pure error' or within-covariate combination variability. In this approach, $\boldsymbol{\gamma}$ can be estimated by solving

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{w_{ij}(y_{ij} - \bar{y}_{wi})^2}{v(\bar{y}_{wi}; \boldsymbol{\gamma})} = \sum_{i=1}^{n} (n_i - 1).$$

With such a estimator for $\boldsymbol{\gamma}$, S still has meaning as a goodness of fit statistic.

6 Unknown Dispersion Parameter

All the above results have assumed that $\phi = 1$. If ϕ is unknown, then both $\dot{\ell}$ and $\mathcal{I}$ are divided by ϕ and the score statistic for goodness of fit for a generalized nonlinear model becomes

$$S = \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{\phi v(\hat{\mu}_i)}.$$

The appearance of the unknown scale parameter ϕ in S means that the statistic is no longer useful for judging goodness of fit. The statistic leads instead, by equating S to its expectation, to the so-called Pearson estimator of ϕ,

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}$$

which is the default estimator of ϕ in generalized linear model functions in the statistical programs S-Plus and R. Other estimators of ϕ are discussed by Jørgensen [11].

When there are repeat observations, the difference between the full version of the score statistic in Theorem 1 and the reduced form in Corollary 1 can be used to define a 'pure error' estimate of the dispersion parameter ϕ,

$$\tilde{\phi}_{\text{pure}} = \frac{1}{\sum(n_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{w_{ij}(y_{ij} - \bar{y}_{wi})^2}{v(\bar{y}_{wi})}.$$

In the case of normal linear regression, this is the well known 'pure error' estimator of the variance. With the use of this this estimator, the score statistic recovers its use as a goodness of fit statistic, but now as a generalized F-statistic rather than chisquare. Substituting the pure error estimator into the score test for the reduced data gives

$$F = \frac{\sum(n_i - 1)}{n-p} \sum_{i=1}^{n} \frac{w_{i\cdot}(\bar{y}_{wi} - \hat{\mu}_i)^2}{\tilde{\phi}_{\text{pure}} v(\hat{\mu}_i)}.$$

If the y_{ij} are approximately normal, then F follows approximately an F-distribution on $n-p$ and $\sum(n_i - 1)$ degrees of freedom under the null hypothesis. This is asymptotically true for example as the weights $w_{ij} \to \infty$ or the dispersion $\phi \to 0$, because any exponential dispersion model $\text{ED}(\mu, \phi)$ tends to normality as $\phi \to 0$ [11, 12]. The F statistic above is a generalization of the normal theory equivalents, described for example by Weisberg [28, Section 4.3].

Dedication

This article is in honor of Terry Speed, from whom I learned generalized linear models while an undergraduate student in Perth, Western Australia. Terry's enthusiasm for statistics and science was and remains infectious. The topic of this article partly arises from a more recent conversation with Terry.

Gordon K. Smyth, Division of Genetics and Bioinformatics, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, smyth@wehi.edu.au.

References

[1] Breslow, N. E. Score tests in overdispersed generalized linear models. In A. Decarli, B. J. Francis, R. Gilchrist, and G. U. H. Seeber, editors, *Proceedings of GLIM 89 and the 4th International Workshop on Statistical Modelling*, pages 64–74. Springer-Verlag: New York, 1989.

[2] Breslow, N. E. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, 85:565–571, 1990.

[3] Cameron, A., and Trivedi, P. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46:347–364, 1990.

[4] Chen, C.-F. Score tests for regression models. *Journal of the American Statistical Association*, 78:158–161, 1983.

[5] Cox, D. R. Some remarks on overdispersion. *Biometrika*, 70:269–274, 1983.

[6] Cox, D. R., and Hinkley, D. V. *Theoretical Statistics.* Chapman and Hall: London, 1974.

[7] Dean, C. B. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87:451–457, 1992.

[8] Deng, D., and Paul, S. R. Score tests for zero inflation in generalized linear models. *Canadian Journal of Statistics*, 28:563–570, 2000.

[9] Genter, F. C. and Farewell, V. T. Goodness-of-link testing in ordinal regression models. *Canadian Journal of Statistics*, 13:37–44, 1985.

[10] Jørgensen, B. Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society Series* B, 49:127–162, 1987.

[11] Jørgensen, B. The theory of exponential dispersion models and analysis of deviance. Monografias de Matemátika No. 51, Instituto de Mathemátika pura e Aplicada, Rio de Janeiro, 1992.

[12] Jørgensen, B. *Theory of Dispersion Models.* Chapman and Hall: London, 1997.

[13] Lu, W.-S. Score tests for overdispersion in Poisson regression models. *Journal of Statistical Computation and Simulation*, 56:213–228, 1997.

[14] McCullagh, P., and Nelder, J. A. *Generalized Linear Models.* Chapman and Hall: London, 1989.

[15] Neyman, J. Optimal asymptotic tests of composite hypotheses. In V. Grenander, editor, *Probability and Statistics: The Harold Cramér Volume*, pages 213–234. Wiley: New York, 1959.

[16] Neyman, J., and Scott, E. On the use of $C(\alpha)$ optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute, Proceedings of the 35th Session*, 41:477–497, 1966.

[17] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50:157–175, 1900.

[18] Paul, S. R., and Deng, D. Goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society Series* B, 62:323–333, 2000.

[19] Paul, S. R., and Banerjee, T. Analysis of two-way layout of count data involving multiple counts in each cell. *Journal of the American Statistical Association*, 93:1419–1429, 1998.

[20] Pregibon, D. Goodness of link tests for generalized linear models. *Applied Statistics*, 29:15–24, 1980.

[21] Pregibon, D. Score tests in GLIM with applications. In R. Gilchrist, editor, *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, pages 87–97. Springer-Verlag: New York, 1982.

[22] Rao, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44:50–57, 1947.

[23] Rao, C. R. *Linear Statistical Inference and its Applications*, Second Edition. Wiley: New York, 1973.

[24] Smith, P. J., and Heitjan, D. F. Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics*, 42:31-41, 1993.

[25] Smyth, G.K. Exponential dispersion models and the Gauss-Newton algorithm. *Australian Journal of Statistics*, 33:57–64, 1991.

[26] Thall, P. F. Score tests in two-way layouts of counts. *Communications in Statistics Part A—Theory and Methods*, 21:3017–3036, 1992.

[27] Wei, B.-C. *Exponential Family Nonlinear Models.* Springer-Verlag: Singapore, 1998.

[28] Weisberg, S. *Applied linear regression.* Wiley: New York, 1985.