Farrington's modified Pearson statistic

In generalized linear models* (GLMs) the Pearson statistic is sometimes applied for assessing the goodness of fit* or to estimate an additional dispersion parameter in the variance model. In case of goodness of fit its conditional distribution given the sufficient statistic* for the regression parameters is of importance (see McCullagh [5]). Approximative conditional and unconditional moments are derived in McCullagh [3] for linear exponential family* models where the natural (or canonical) parameter is modeled linearly in some explanatory variables. In the context of a GLM this situation corresponds to a so-called *canonical link model.* To extend these results to arbitrary link functions a family of first-order modifications of Pearson's statistic is introduced in Farrington [2]. Limits are considered for sparse data where the residual degrees of freedom* (d.f.) becomes large.

The null model. Consider responses y_i, $i = 1, \ldots, n$, sampled independently from an exponential family distribution with means μ_i. Hence, $var(y_i) = V_i = \kappa_{2i}$ and all higher order cumulants* $\kappa_{k+1,i} = \kappa'_{ki} V_i$, $k \geq 2$, are functions of μ_i. In the following we assume that there is structure in the means which can be described by the GLM

$$h^{-1}(\mu) = \eta = X\beta. \tag{1}$$

h denotes an inverse link function, X is a $n \times p$ model matrix and β is the $p \times 1$ vector of unknown parameters with maximum likelihood* estimator $\hat{\beta}$. Note that $h'_i = dh(\eta_i)/d\eta_i = V_i$ and $h''_i = V'_i V_i = \kappa_{3i}$, with $V'_i = dV_i/d\mu_i$, holds for canonical link models.

Goodness of fit. To assess goodness of fit, the model is embedded in a wider family, again with $E(y_i) = \mu_i$ and model (1) but with variance ϕV_i and cumulants $\phi^{k-1}\kappa_{ki}$, which are proportional to those under the null hypothesis where $\phi = 1$. This covers the problem of *overdispersion* in Poisson* or binomial* models, where the variability of the residuals is systematically larger than specified by the model. Departures from the null model are commonly evaluated by the (generalized) *Pearson statistic*

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{V}_i}. \tag{2}$$

Farrington [2] discusses the family

$$X_a^2 = X^2 + \sum_{i=1}^n \hat{a}_i (y_i - \hat{\mu}_i) \tag{3}$$

which includes (2) by setting $a_i = 0$. Generally, the a_i are functions of μ_i not depending on ϕ. Because of $cov(X_a^2, \hat{\beta}) = O(n^{-1})$, the regression parameter estimates and the family (3) are asymptotically uncorrelated.

Local orthogonality between $\hat{\beta}$ and X_a^2 is induced under the null model only if $a_i = -V_i'/V_i$. Hence, the *Farrington statistic*

$$X_*^2 = X^2 - \sum_{i=1}^n \frac{\hat{V}_i'}{\hat{V}_i}(y_i - \hat{\mu}_i) \tag{4}$$

depends only weakly on β, given $\hat{\beta}$, and can be used to assess goodness of fit. If $\phi = 1$, the special choice $a_i = -V_i'/V_i$ eliminates the dependence of $E(X_a^2)$ from the first-order bias* of $\hat{\beta}$, minimizes the variance under all members of the family (3), i.e. $var(X_*^2) \leq var(X_a^2)$, and results in much simpler expressions for the first three moments. If ϕ is treated as unknown then none of these properties generally continue to hold. However, under the null model it follows that

$$\begin{aligned}
E(X_*^2|\hat{\beta}) &= n - p - \frac{1}{2}\sum_{i=1}^n \frac{\hat{V}_i''}{\hat{V}_i}\hat{h}_i'^2 \hat{Q}_{ii} + O(n^{-1/2}) \\
var(X_*^2|\hat{\beta}) &= (n-p)\left(\hat{\rho}_4 - \hat{\rho}_3^2 + 2\right) + O(n^{1/2}) \\
\kappa_3(X_*^2|\hat{\beta}) &= (n-p)(\hat{\rho}_6 - 3\hat{\rho}_{35} + 3\hat{\rho}_{34}^2 + 12\hat{\rho}_4 - 8\hat{\rho}_3^2 - \hat{\rho}_3^4 + 8) + O(n^{1/2})
\end{aligned} \tag{5}$$

with standardized cumulants

$$\hat{\rho}_k^s = n^{-1}\sum_i \hat{\kappa}_{ki}^s / \hat{V}_i^{sk/2}, \qquad \hat{\rho}_{kl}^s = n^{-1}\sum_i \hat{\kappa}_{ki}^s \hat{\kappa}_{li} / \hat{V}_i^{(sk+l)/2}$$

and the matrix $\hat{Q} = X(X^t \hat{W} X)^{-1} X^t$, where $\hat{W} = diag(\hat{h}_i'^2/\hat{V}_i)$, which estimates the covariance matrix of the fitted linear predictor $\hat{\eta} = X\hat{\beta}$. For some relevant models the conditional moments (5) may be found in Farrington [1].

An approximate significance level is calculated by referring

$$Z = (X_*^2 - E(X_*^2|\hat{\beta}))/var^{1/2}(X_*^2|\hat{\beta})$$

to the standard normal distribution or by the Edgeworth approximation (*see* CORNISH-FISHER AND EDGEWORTH EXPANSIONS)

$$P(Z \geq z|\hat{\beta}) = 1 - \Phi(z) + \phi(z)(z^2 - 1)\rho_3/6 \tag{6}$$

where ρ_3 is the standardized conditional skewness of Z.

Example 1. To illustrate the effect of using X_*^2, a logistic regression* model is considered for data on the recurrence of cervical carcinoma given in Table 1. Hence, we assume that $y_i \sim Bin(m_i, \pi_i)$ with $\mu_i = m_i\pi_i = h(\eta_i) = \exp(\eta_i)/(1 + \exp(\eta_i))$ and $V_i = \mu_i(m_i - \mu_i)/m_i$. Tumor prognosis is modeled by a tumor extension factor E (1-cervix only, 2-border zone involved, 3-parametria involved) and the number L of lymph node stations involved. A main effects model results in $X^2 = 11.49$ on 6 d.f. with conditional moments 6.302, 12.13 and 1.542 by using the macro in McCullagh [5]. The respective Edgeworth approximation gives an upper tail probability of $p = 0.109$. Farrington's $X_*^2 = 10.14$ is based on $a_i = -(m_i - 2\mu_i)/\mu_i(m_i - \mu_i)$ and the approximate conditional moments

$$\begin{aligned}
E(X_*^2|\hat{\beta}) &= n - p + \sum_{i=1}^{n} \hat{V}_i\hat{Q}_{ii}/m_i \\
var(X_*^2|\hat{\beta}) &= 2(n-p)\frac{1}{n}\sum_i (m_i - 1)/m_i \\
\kappa_3(X_*^2|\hat{\beta}) &= 8(n-p)\left(1 - \frac{1}{n}\sum_i (5m_i - 4)/m_i^2 + \frac{1}{2n}\sum_i (m_i - 1)/m_iV_i\right)
\end{aligned}$$

are 6.357, 11.07 and 1.314. From (6) we get $p = 0.141$ showing slightly less evidence of overdispersion relative to the binomial variability. Note that in case of binary responses ($m_i = 1$), we have $X_*^2 = n$ which is completely uninformative.

Table 1: Recurrence rates y/m from cervical carcinoma

	L			
E	0	1	2	$\geq$3
1	21/124	7/21	9/16	13/13
2	18/ 58	6/12	5/ 7	5/ 5
3	4/ 14	16/19	9/12	10/12

Example 2. Since the d.f. in Example 1 is very small we also analyze the data from the Ille-et-Vilaine study as in McCullagh [3]. Pearson's X^2 is 86.56 on 76 d.f. with conditional moments 77.38, 401.1 and 1.985 corresponding to $p = 0.230$. The advantage of Farrington's X_*^2, which is 84.40 with moments 78.03, 108.3 and 1.295, is its smaller variance and third cumulant. Although the Edgeworth expansion (6) results in a similar p-value of 0.226, this approximation should be more accurate.

References

[1] Farrington, C. P. (1995). In *Statistical Modelling*, G.U.H. Seeber, B.J. Francis et al., eds., Springer, New York, pp. 109–116.

[2] Farrington, C. P. (1996). On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data. *J. R. Statist. Soc. Ser. B*, **58**, 349–360.

[3] McCullagh, P. (1985). On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models. *Int. Statist. Rev.*, **53**, 61–67.

[4] McCullagh, P. (1985). Macro to Calculate the Approximate Conditional Cumulants of Pearson's Goodness-of-Fit Statistic for Binomial and Poisson Response Models. *Glim Newsletter*, **10**, 41–42.

[5] McCullagh, P. (1986). The Conditional Distribution of Goodness-of-Fit Statistics for Discrete Data. *J. Amer. Statist. Ass.*, **81**, 104–107.

(CHI-SQUARE TESTS
GENERALIZED LINEAR MODELS
GOODNESS OF FIT
LOGISTIC REGRESSION)