Continuity Correction for the Score Statistic in Discrete Regression Models

By John E. Kolassa[1]

Rutgers University Department of Statistics

This paper corrects the usual chi–squared approximation to the distribution function of the conditional score statistic in a generalized linear model, when the underlying distribution is discrete. The proposed method corrects by a multiple of the difference between the number of sufficient statistics lying in the acceptance region for the test and the volume of this region. The multiplier is calculated from the multivariate Edgeworth approximation to the distribution of a lattice random vector.

1. Introduction. This paper addresses the problem of hypothesis testing in generalized linear models in the presence of a canonical nuisance parameter. Marginal approaches involving the score statistic are fully efficient [8], but have the drawback that the sampling distribution of the test statistic depends, at least weakly, on the nuisance parameter. Conditional inference avoids problems arising from this dependence, often at a cost in efficiency that is not particularly severe [7]. This paper applies a continuity correction to the standard χ^2 approximation to the distribution of the conditional score statistic.

When the distribution of raw responses is continuous, standard Edgeworth techniques may be employed to improve on the usual normal theory approximation to the test statistic sampling distribution. When the distribution of raw responses is discrete, standard Edgeworth series results do not apply. This paper applies a continuity correction to the estimation of probabilities associated with the conditional scores statistic arising in generalized linear models. Approximations to the conditional expectation and variance, as presented by Waterman and Lindsay [18], are used in conjunction with a first Edgeworth correction term calculated by Yarnold [17], to accurately approximate p-values.

Section 2 reviews the multivariate Edgeworth series, and section 3 discusses estimation of probabilities for ellipses that arise from score testing, and reviews an adjustment to standard approximations that accounts for the lattice nature of certain regression models. Section 4 reviews generalized linear model notation. Section 5 presents an artificial multinomial example, and section 6 presents an example concerning cancer remissions.

2. Multivariate Edgeworth Series. Suppose $\boldsymbol{X} = (X^r)$ is a random vector in $\boldsymbol{R}^m$, such that

$$\mathrm{E}[\boldsymbol{X}] = \boldsymbol{o} \text{ and } \mathrm{Var}[\boldsymbol{X}] = \boldsymbol{\Sigma}. \tag{1}$$

[1]This research was supported by grant R01 CA 63050 from the National Cancer Institute.