TI 2001-082/4
Tinbergen Institute Discussion Paper

Measures of Fit for Multinomial Discrete Models

J.S. Cramer

Tinbergen Institute

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

Measures of Fit for Multinomial Discrete Models

J.S. Cramer*

August 28, 2001

Abstract

In a discrete model, the predicted probabilities of a particular event can be matched to the observed $(0, 1)$ outcomes and this will give rise to a measure of fit for that event. Previous results for the binomial model are applied to multinomial models. In these models the measure of fit will vary between the various events, indicating that the model performs better for some than for others. In addition to these differential measures of fit for each separate event a single overall measure is constructed for the model as a whole.

Key Words & Phrases: Goodness of fit, orthogonal residuals, R^2, multinomial logit model.

1 Introduction and summary

A multinomial discrete choice model describes the occurrence of S alternative states or events. An assessment of the goodness of fit for one event, considered in isolation, is obtained by matching predicted and observed outcomes, where the former are probabilities and the observations are $0, 1$ variables. This can be done in various ways, some of which are listed in the survey by Menard (2000). The issue was treated in an earlier paper (Cramer (1999)) for binary

*Tinbergen Institute, Keizersgracht 482, 1017 EG Amsterdam, Holland; e-mail mars.cram@worldonline.nl. I am greatly indebted to Jan Sandee and Geert Ridder for useful comments on an earlier version.

models, and the argument that led to the adoption of λ (which comes close to R^2) as a measure of fit is repeated in Section 2. In a binary model this statistic (like other measures of fit) will have the same value for an event and its complement. With more than two states, however, the results for the S alternatives generally differ. This indicates that the model fits some events better than others, and thus suggests where there is room for improvement. In addition to these differential measures of fit, a single overall measure for the model as a whole is in order. These aspects of the multinomial case are the subject of the present paper. The favoured measure for a single event turns out to be asymptotically identical (and hence a close approximation) to R^2, the classical coefficient of determination, for a OLS regression of discrete outcomes on the corresponding probabilities. Its use in discrete models is not novel. To my knowledge, however, the derivation and supporting theory of the 1999 paper was new, and it is this that is now generalized to the multinomial case.

The argument makes use of asymptotic properties of the predicted probabilities that are established in another companion paper (Cramer (2000)). Whether these hold approximately for a finite sample is easily verified; they usually do. With this reservation, the present results apply to any well behaved probabilistic model of a discrete outcome, and not just to the logistic regressions that serve as illustrations. But as the asymptotic properties have been derived for random samples from a given population, they are more appropriate to survey data from the social sciences and epidemiology than to controlled experiments.

Section 2 recalls the main points from Cramer (1999), Section 3 addresses the general case of S alternatives, considered jointly, and Section 4 gives two illustrations from multinomial logit analyses.

2 Closeness of fit for a single event

A multinomial model determines the conditional probabilities of S discrete events at observation i as a function of known regressor variables x_i. In this section we consider a single event or state s in isolation, and temporarily omit the suffix s. $Y_i = 1$ denotes the event, and the model gives its probability

$$P_i = P(Y_i = 1|x_i) = P(x_i, \theta).$$

Consistent parameter estimates $\hat{\theta}$ (like the standard Maximum Likelihood estimates) are obtained from a random sample of size n from a given population, and these yield *predicted probabilities*

$$\hat{P}_i = P(x_i, \hat{\theta}).$$

The sample outcomes and predicted probabilities are arranged in vectors y and $\hat{p}$. Since $E(y) = p$ it is natural to define *crude residuals*

$$e = y - \hat{p}.$$

These residuals share two properties of the ordinary least squares residuals of linear regression, if only asymptotically. The first is the *zero mean property*

$$\imath^T e/n \xrightarrow{p} 0, \tag{1}$$

where $\imath$ denotes a vector of ones. In terms of $\hat{p}$ and y this implies

$$\imath^T \hat{p} \xrightarrow{p} \imath^T y. \tag{2}$$

The second property is *orthogonality*

$$\hat{p}^T e/n \xrightarrow{p} 0. \tag{3}$$

For samples of reasonable size these asymptotic properties will hold approximately. The first leads from (2) to the *equality of means:* the mean sample probability is (approximately) equal to the sample incidence or base rate α,

$$\bar{p} = \imath^T \hat{p}/n \approx \imath^T y/n = \alpha. \tag{4}$$

In the special case of a logit model this holds exactly. - Similarly, (3) leads to the approximation

$$\hat{p}^T e = \hat{p}^T (y - \hat{p}) \approx 0. \tag{5}$$

Since e has zero mean (if only approximately) it follows at once that e and $\hat{p}$ are approximately uncorrelated. And (5) also implies

$$\hat{p}^T y \approx \hat{p}^T \hat{p}. \tag{6}$$

A simple and direct way to assess the fit is to examine the estimated probabilities of the observations with $Y_i = 1$. The higher these probabilities, the lower the probabilities of the other observations, since the overall average is constrained by (4); and the better is the within-sample predictive performance of the model. The mean of the probabilities for $Y_i = 1$ is

$$\bar{P}^+ = \hat{p}^T y / n_1, \tag{7}$$

with $n_1 = \alpha.n$ the frequency of $Y_i = 1$. As in all measures of fit (beginning with the R^2 of classic linear regression) we define two limiting cases for this quantity and ascertain its actual position in the intervening interval. Here the bottom line is the performance of the *null model*, with a single constant as the sole regressor, or zero coefficients for all (other) regressor variables. In this case the same constant probabilities apply to all observations, regardless of the outcome Y_i; in accordance with (4) these are equal to α, and so is the minimum $\bar{P}^+$. The upper limit obtains for the near-perfect model with the $\hat{P}_i^+$, and hence $\bar{P}^+$, arbitrarily close to 1. The actual position of $\bar{P}^+$ can be represented by writing it as a weighted average of these bounds with weights λ and $1 - \lambda$, $0 \leq \lambda < 1$,

$$\bar{P}^+ = \lambda + (1 - \lambda)\alpha, \tag{8}$$

or

$$\lambda = \frac{\bar{P}^+ - \alpha}{1 - \alpha}. \tag{9}$$

λ thus indicates the position of $\bar{P}^+$ between α and 1 on a $(0, 1)$ scale.

There are several other equivalent expressions for λ. A counterpart to $\bar{P}^+$ is $\bar{P}^-$, the mean probability over the other observations, with $Y_i = 0$; as it turns out,

$$\lambda = \bar{P}^+ - \bar{P}^-,$$

so that λ measures the discrimination of P_i between observations with and without the event under consideration.

By (3) $\hat{P}_i$ and e_i are asymptotically uncorrelated, and they will be nearly so in samples of reasonable size. The sum of squares of Y_i can therefore be decomposed as

$$SS_y \approx SS_p + SS_e. \tag{10}$$

By (6), (7) and (8) SS_p may be expressed in n, α and λ; SSS_y is equal to $n\alpha(1-\alpha)$. Substituting these values we find

$$\lambda \approx SS_p/SS_y \approx 1 - SS_e/SS_y, \tag{11}$$

so that λ is a close approximation of the R^2 of a linear regression of y and $\hat{p}$. This result depends essentially on the orthogonality of $\hat{p}$ and e, which follows from the asymptotic properties (1) and (3).

The use of this R^2 as a measure of fit for discrete models has been suggested before by Efron (1979), Maddala (1983, pp.38-39) and Agresti (1996, p.129); as R^2_{OLS} it is the first on the list of Menard (2000). Most of these authors justify the measure by mere analogy (if at all); Efron is an exception in that he takes great trouble over the decomposition of the sum of squares of y along the lines of (10). As he considers a more detailed decomposition in several successive components he has recourse to a fairly special sample design to ensure their orthogonality.

3 The multinomial case

In a multinomial model outcomes and probabilities are recorded in $n \times S$ matrices Y and $\hat{P}$ with columns y_s and p_s. Upon adding the suffix s all results of the preceding section apply to each event s considered separately. The only exception is the generalization of (3); it turns out that the $\hat{p}_s$ are asymptotically uncorrelated with *all* residuals e_t,

$$\hat{p}_s^T e_t/n \xrightarrow{p} 0 \quad \forall s,t.$$

Hence

$$\hat{P}^T Y \approx \hat{P}^T \hat{P}, \tag{12}$$

which goes further than (6); it implies symmetry of the matrix $\hat{P}^T Y$ which has the sums of probabilities over observations with a given outcome as its elements.

For each s, a separate λ_s can be derived from the columns $\hat{p}_s$ and y_s in the same manner as before; take $\hat{p}_s^T y_s$, a diagonal element of $\hat{P}^T Y$, divide by the frequency n_s to obtain the mean probability $\bar{P}_s^+$, and find λ_s as in (9),

$$\lambda_s = \frac{\bar{P}_s^+ - \alpha_s}{1 - \alpha_s}. \tag{13}$$

This will give S values λ_s which reflect the fit for each separate state. As n_s and α_s are given constants from Y, the fitted model intervenes only *via* the diagonal elements of $\hat{P}^T Y$.

To assess the fit of the model as a whole we consider the mean predicted probability of the observed outcome over *all* observations, that is

$$\bar{P}^+ = \sum \alpha_s \bar{P}_s^+. \tag{14}$$

Its minimum is again found by equating the $\bar{P}_s^+$ to their null model values α_s, and this gives $\sum \alpha_s^2$; the upper limit is 1, as before, for a perfect model. The actual position of $\bar{P}^+$ between these bounds is given by the overall $\bar{\lambda}$, as in (8),

$$\bar{P}^+ = \bar{\lambda} + (1 - \bar{\lambda}) \sum \alpha_s^2. \tag{15}$$

As in (9) we have

$$\bar{\lambda} = \frac{\bar{P}^+ - \sum \alpha_s^2}{1 - \sum \alpha_s^2}. \tag{16}$$

Upon rewriting and substituting (13) this gives

$$\bar{\lambda} = \frac{\sum \alpha_s (\bar{P}^+ - \alpha_s)}{1 - \sum \alpha_s^2} = \frac{\sum \alpha_s (1 - \alpha_s) \lambda_s}{\sum \alpha_s (1 - \alpha_s)}. \tag{17}$$

Thus $\bar{\lambda}$ is a weighted average of the state-specific λ_s, with the variances of y_s as weights. It varies between 0 and 1 in the same manner as the λ_s, and if the state-specific λ_s happen to be equal the same value will hold for the overall $\bar{\lambda}$. But $\bar{\lambda}$ does not correspond to a coefficient of determination R^2 for all $n \times S$ observations. By (11) it can be rewritten as

$$\bar{\lambda} \approx \sum (SS_p)_s \ \sum (SS_y)_s,$$

but the numerator and denominator are *not* squared deviations over all observations, for the separate terms SS_p and SS_y are sums of squared deviations from the specific column means.

For $S = 2$ the multinomial model reduces to the binary case. It is easy to see that in that case the same λ applies to the two columns and therefore to the entire model. This is so because y_1, y_2 and $\hat{p}_1, \hat{p}_2$ both sum to $\imath$. If $S > 2$ the columns of $\hat{P}$ and of Y also sum to $\imath$, and this raises the question whether

there are similar restrictions on the λ_s. The answer is negative. The λ_s are *correlated* since the y_s and $\hat{p}_s$ are constrained, but there is no exact side relation among them. The only special cases where one λ_s can be deduced from the $S-1$ others are the two extremes of the null model and of a perfect fit, because for λ_s equal to 0 or to 1 all elements of the corresponding $\hat{p}_s$ are determined.

To see why there are restrictions for $S = 2$, but not for $S > 2$, we recall that the λ_s are derived from the diagonal elements of $\hat{P}^T Y$. The $S \times S$ elements of that matrix are subject to two constraints. First, by the equality of means of (4) the column sums equal the sample frequencies n_s; this constitutes S restrictions. Second, the extended orthogonality condition (12) implies symmetry of $\hat{P}^T Y$, or another $1/2 \times S \times (S-1)$ restrictions. There remain

$$L = 1/2 \times S \times (S-1)$$

elements of $\hat{P}^T Y$ that can vary freely. For $S = 2$, L is 1, and a single λ_s determines the entire matrix. But with $S > 2$ L equals or exceeds S, and there is sufficient room for variation of the λ_s.

4 Two illustrations

Four types of Hepatitis

The first illustration is a multinomial logit analysis of four types of Hepatitis on the basis of a sample of 218 patients by Lesaffre and Albert (1989); one notorious outlier detected by these authors has been removed (nr 136). The data have been collected by Plomteux (1980). The regressors are four types of enzymes. Table 1 gives a number of statistics.

The first thing is to check the approximations that are supposed to hold. Since this is a logit analysis, the equality of the means is automatically satisfied. As for orthogonality, the correlations between the $\hat{p}_s$ and e_t are small, but not very small; but then this is not a large sample.

state:	PNC	AVH	PCH	ACH
$r_{\hat{p}_s, e_t}$	.040 -.004 -.062 .025	-.020 .038 -.006 -.024	-.025 -.024 .057 .002	-.009 -.000 .020 -.013
α_s	.355	.263	.203	.180
$\bar{P}_s^+$	.746	.875	.717	.536
λ_s	.607	.830	.845	.434

Table 1. Statistics for the Hepatitis study

As for the results, the values of λ_s are all very high: on the basis of experience with binary analyses values of .4 and over indicate a strong relationship, and .8 is exceptional. Although there is some variation between the four types, the fit is quite good throughout, and this is reflected in the overall λ of (15) which is .640.

This is an example of an extraordinary good fit, almost uniformly so across all states. It is therefore well suited for a demonstration of the regression diagnostics of Lesaffre and Albert.

Car ownership

The second illustration is a multinomial logit analysis of four classes of car ownership on the basis of a sample survey of some 2800 Dutch households in 1980 that I have analysed elsewhere (Cramer (1991)). There are four classes of car ownership status, viz.

none
one used car
one new car
more cars

and the regressors are the usual determinants like log of income, age, log of family size, urban/rural habitat, and (less usual) a dummy for the presence of a business car.

state:	none	used	new	more
$r_{\hat{p}_s, e_t}$	.006 -.016 -.015 .033	-.006 .012 -.002 -.003	-.004 .009 .015 -.029	.007 -.013 .003 .000
α_s	.358	.335	.245	.062
$\bar{P}_s^+$	.567	.429	.313	.180
λ_s	.325	.141	.090	.125

Table 2. Statistics for the analysis of car ownership

Table 2 gives the statistics. This is a fairly large sample and the correlations of $\hat{p}_s$ and e_t show that the asymptotic properties of section 2 are closely approximated.

The overall $\bar{\lambda}$ is .187, which is fair; the analysis as a whole is satisfactory but certainly not outstanding, and one should look for ways to improve the model. The values of λ_s vary considerably between states; the value of .35 for non-owners is quite high, but .09 for new car ownership is pretty low. Often, the smaller categories are predicted badly, since they pull little weight in the likelihood function that is maximized; but new car ownership represents a quarter of the sample. To improve the overall performance of the model one should concentrate on factors that differentiate between new car ownership and the other three states.

References

Agresti, Alan (1996) *An Introduction to Categorical Data Analysis.* New York: Wiley.

Cramer, J.S. (1991) *The Logit Model: an Introduction for Economists.* London: Arnold.

Cramer, J.S. (1999) Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society, Series D (The Statistician)* **48**, p. 85-94.

Cramer, J.S. (2000) Asymptotic properties of predicted probabilities in discrete regression. *Tinbergen Institute Discussion Paper* 2000-06/4. *available from www.tinbergen.nl.*

Efron, Bradley (1978) Regression and ANOVA with zero-one data: measures of residual variation. *Journal of the American Statistical Association* **73**, 113-121.

Lesaffre, E., and A. Albert (1989) Multiple-group Logistic Regression Diagnostics. *Applied Statistics* **38**, 425-445.

Maddala, G.S. (1983) *Limited-dependent and qualitative variables in econometrics.* Cambridge: Cambridge University Press.

Menard, Scott (2000) Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician* **54**, 17-24.

Plomteux, G. (1980) Multivariate analysis of enzyme profile for the differential diagnosis of viral hepatitis. *Clinical Chemistry* **26**, 1897-1899.