TI 2000-06/4
Tinbergen Institute Discussion Paper

Asymptotic Properties of Predicted Probabilities in Discrete Regression

J.S. Cramer[1]

[1] *University of Amsterdam and 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

Asymptotic Properties of Predicted Probabilities in Discrete Regression

J.S. Cramer *

december 1999

Abstract

The discrete outcome of a probability model is recorded as $Y_i = 1$ while otherwise $Y_i = 0$. y is the vector of observed outcomes, p the corresponding probabilities, $\hat{p}$ a consistent estimate of p, and residuals are defined as $e = y - \hat{p}$. Under quite general conditions, the asymptotic properties of $\hat{p}$ ensure that these residuals have zero mean and are uncorrelated with $\hat{p}$. These asymptotic results extend to the multinomial case. They support certain measures of fit for discrete models.

1 Introduction

Two well known properties of residuals of common linear regression are that they have zero sample mean and are orthogonal to the regressor variables, and hence to the predicted values of the dependent variable. This paper establishes closely similar asymptotic properties for discrete probability models. The result is used in two other papers to justify a measure of fit: for the binary model in Cramer (1999), for multinomial models in Cramer (2000). Both papers provide empirical illustrations of the properties under review.

*Tinbergen Institute, Keizersgracht 482, 1017 EG Amsterdam, Holland; e-mail Cramer@Tinbinst.nl. This is a revised version of discussion paper TI 97-044/4. I am greatly indebted to Geert Ridder and Peter Boswijk for guidance and support.

Consider a single discrete variate that obeys a given probability regression model with parameter vector θ. The model defines the probability that the outcome Y_i is 1 at regressor vector x_i as

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

Consistent parameter estimates $\hat{\theta}$ have been obtained from a random sample of size n from a given population, for example by Maximum Likelihood estimation. These estimates determine *predicted probabilities*

$$\hat{P}_i = P(x_i, \hat{\theta}). \tag{2}$$

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

$$e = y - \hat{p}. \tag{3}$$

These differ from the common linear regression residuals, but they do share two major properties, if only asymptotically. First, with $\imath$ the unit vector, the *zero mean property*

$$\imath^T e/n \xrightarrow{p} 0. \tag{4}$$

In terms of $\hat{p}$ and y this implies the *equality of means;* by

$$\imath^T \hat{p} \approx \imath^T y$$

the mean of the $\hat{P}_i$ is equal to the sample frequency α of the event $Y_i = 1$,

$$\bar{p} = \imath^T \hat{p}/n \approx \imath^T y/n = \alpha. \tag{5}$$

Second, the *orthogonality property* is

$$\hat{p}^T e/n \xrightarrow{p} 0. \tag{6}$$

Since e has zero mean, this means that e and $\hat{p}$ are approximately uncorrelated. In terms of $\hat{p}$ and y (6) gives

$$\hat{p}^T \hat{p} \approx \hat{p}^T y. \tag{7}$$

We establish these properties below. The major assumption is that the pairs (x_i, Y_i) are a random sample from a given population, which is more appropriate for analyses of survey data than for controlled experiments. But there are no restrictions on the context of the single outcome Y_i: it may belong to a narrow binary model, where the only alternative is its complement, or it may be part of a larger or more complex model. The properties therefore hold for each outcome of a multinomial discrete model and also for the discrete arm of a mixed model like the Tobit model.

2 Practical relevance

Two issues determine the practical relevance of these asymptotic results, namely whether they are approximately valid in finite samples, and whether they are of any use.

The first question is easily settled, for it is readily verified in any particular instance whether the sample mean residual $\bar{e}$ and the sample correlation $r(\hat{p}, e)$ are close to zero, as they should be by (4) and (6). In a bivariate logit model (4) holds *exactly*, for the Maximum Likelihood estimates $\hat{p}_l$ satisfy

$$X^T(y - \hat{p}_l) = X^T e = 0, \tag{8}$$

and (4) follows if the regressor matrix X contains a column of constants (as it usually does). As for (6), the logit probabilities $\hat{p}_l$ are often quite close to a linear function of X, and then (8) ensures that $r(\hat{p}, e)$ is close to zero.

So far the usefulness of the properties is that they support a measure of fit since they imply that the classical decomposition of sums of squares is approximately valid for discrete outcomes. From (6) we have

$$y^T y \approx \hat{p}^T \hat{p} + e^T e. \tag{9}$$

By the zero mean property, y and $\hat{p}$ have a common mean, and upon taking this out (9) can be rewritten in terms of sums of squares around the mean as

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

The observed variation of y is thus split up into two orthogonal components, and the sum of squares of y is decomposed into an explained part and an unexplained residual part. This immediately suggests

$$\lambda = 1 - SS_e/SS_y \tag{11}$$

as a measure of fit. This is identical to the R^2 proposed by Efron (1978), who ensures that (10) holds by considering the special case of grouped data. This has restricted the general adoption of his measure, as noted by Windmeijer (1996). The present result vindicates a much wider use of R^2 or λ.

3 Assumptions

We consider a random sample of n independent observations (x_i, Y_i), drawn from a given population, and consistent estimates $\hat{p}$ of the P_i of (1). Consistency implies

$$\hat{P}_i \xrightarrow{p} P_i. \tag{12}$$

The $\hat{P}_i$ are not independent, for they share $\hat{\theta}$ and satisfy side restrictions like (8).

Amemiya (1985) has shown that for the models under consideration Maximum Likelihood estimation provides consistent parameter estimates. This is the usual procedure. The parameter space Θ is then defined in such a way that P_i is a proper probability for any x_i and any admissible θ and $\hat{\theta}$. The P_i and $\hat{P}_i$ are thus restricted to the open interval $(0, 1)$; its bounds are excluded since the model (as well as ML estimation) breaks down at these limiting values. The argument that follows makes use of the consistency (12) and of the restriction of Y_i, P_i and $\hat{P}_i$ to the interval $(0, 1)$. Moreover, the treatment of observed pairs (x_i, Y_i) as independent random drawings from a given population ensures that *all* statistics for a given observation, including $\hat{P}_i$, have the same distribution for all i.

These assumptions are met in a wide range of cases. The restrictions on the parameter space are a standard matter and consistency is a common property of estimates. The treatment of the (x_i, Y_i) as independent random drawings from a given population may not be appropriate for controlled experiments, but it does apply to survey data, under the pretense that the observations have been drawn with replacement. The argument is thus valid for analyses of such data in epidemiology and the social sciences.

4 A lemma

For the record we establish the following result: Let

$$S^{(n)} = 1/n \sum Z_i^{(n)} \tag{13}$$

where the $Z_i^{(n)}$ are identically distributed, restricted to the interval $(0, 1)$ and have probability limit zero,

$$Z_i^{(n)} \xrightarrow{p} 0. \tag{14}$$

Under these conditions

$$S^{(n)} \xrightarrow{p} 0. \tag{15}$$

The proof starts from the probability limit (14). By its definition there is for any η and δ, however small, a n^* such that for all $n > n^*$

$$P\left(Z_i^{(n)} > \eta\right) < \delta. \tag{16}$$

Note that, for given η and δ, the same n^* applies to all $Z_i^{(n)}$ because they have the same distribution. In the sequel n is always taken to exceed n^* so that (16) holds.

Now consider the expected value of $Z_i^{(n)}$. With a probability density $f(z)$ this is

$$E\ Z_i^{(n)} = \int_0^1 zf(z)dz = \int_0^\eta zf(z)dz + \int_\eta^1 zf(z)dz.$$

Obviously

$$\int_0^\eta zf(z)dz \leq \eta$$

and, by (16),

$$\int_\eta^1 zf(z)dz \leq \delta.$$

As a result

$$E\ Z_i^{(n)} \leq \eta + \delta$$

and hence

$$E\ S^{(n)} = E\ 1/n \sum Z_i^{(n)} \leq \eta + \delta. \tag{17}$$

Since the $Z_i^{(n)}$ are nonnegative so is $S^{(n)}$ and Markov's inequality applies, or

$$P(|S^{(n)}| \geq \epsilon) \leq \frac{E|S^{(n)}|}{\epsilon},$$

or, with (17),

$$P(S^{(n)} \geq \epsilon) \leq \frac{\eta + \delta}{\epsilon} = \delta^*.$$

Hence

$$S^{(n)} \xrightarrow{p} 0,$$

which is the desired result.

This lemma will be applied to functions of Y_i, P_i and $\hat{P}_i$; for the latter estimates each $Z_i^{(n)}$ depends on the entire sample, and the upper index (n) recalls this fact. Since all samples are drawn in the same manner from the same population, the $Z_i^{(n)}$ are identically distributed; but as the $\hat{P}_i$ depend on a single set of parameter estimates, the $Z_i^{(n)}$ are *not* always independent.

5 Zero mean residuals

To show the zero mean property (1)

$$\mu = 1/n \sum (Y_i - \hat{P}_i) \xrightarrow{p} 0$$

rewrite μ as

$$\mu = 1/n \sum (Y_i - P_i) - 1/n \sum (\hat{P}_i - P_i) \tag{18}$$

and take the two terms in turn.

First define

$$E\, Y_i = E_{x_i}\, E\, (Y_i | x_i) = E\, P_i = EP.$$

By the law of large numbers

$$1/n \sum Y_i \xrightarrow{p} EP, \quad 1/n \sum P_i \xrightarrow{p} EP$$

so that

$$1/n \sum (Y_i - P_i) \xrightarrow{p} 0.$$

This takes care of the first term.

For the second term first write

$$1/n \sum (\hat{P}_i - P_i) \leq 1/n \sum |(\hat{P}_i - P_i)|$$

By the consistency condition (12) each element of the summation converges to zero, and as the absolute values are moreover constrained to the interval (0,1) the lemma of section 4 applies; consequently

$$1/n \sum (\hat{P}_i - P_i) \xrightarrow{p} 0.$$

As both terms converge to zero, so does their difference, and this establishes (18).

6 Orthogonality

To show this property (2) or

$$\nu = 1/n \sum \hat{P}_i(Y_i - \hat{P}_i). \tag{19}$$

is rewritten in four terms as

$$\nu = 1/n \sum (\hat{P}_i - P_i)Y_i + 1/n \sum P_i Y_i - 1/n \sum (\hat{P}_1^2 - P_i^2) - 1/n \sum P_i^2.$$

For the first term

$$1/n \sum (\hat{P}_i - P_i)Y_i \leq 1/n \sum |(\hat{P}_i - P_i)Y_i)| \xrightarrow{p} 0$$

since by (12) the conditions of the lemma apply.

For the second term

$$1/n \sum P_i Y_i \xrightarrow{p} E\ P_i Y_i = E\ P_i E\ (Y_i | P_i) = EP^2.$$

For the third term

$$1/n \sum (\hat{P}_i^2 - P_i^2) \leq 1/n \sum |(\hat{P}_i^2 - P_i^2)|$$

and

$$1/n \sum |(\hat{P}_i^2 - P_i^2)| \xrightarrow{p} 0$$

since by Slutsky's theorem and (12) the conditions of the lemma apply.

The fourth term at once satisfies

$$1/n \sum P_i^2 \xrightarrow{p} EP^2$$

Collecting the four terms yields

$$\nu \xrightarrow{p} 0 \tag{20}$$

which is the orthogonality property.

7 Multinomial Extension

So far the vectors $\hat{p}$, y and e refer to a single event. This may well be a particular alternative s among a set of S alternatives in a multinomial model. The vectors then denote corresponding columns $\hat{p}_s$, y_s and e_s of $S \times n$ matrices $\hat{P}$, Y and E.

The zero mean property carries over at once to this larger set and reads

$$\imath^T E/n \xrightarrow{p} 0, \tag{21}$$

with the right-hand 0 now a $Sx1$ vector. For the sample matrix $\hat{P}$ this implies

$$\imath^T \hat{P} \approx \imath^T Y, \tag{22}$$

that is the *equality of means.*

The generalization of the orthogonality property is not so trivial and yields a new result. As long as we consider a single event, (2) is interpreted naturally as

$$\hat{p}_s^T e_s/n \xrightarrow{p} 0$$

and (19) as

$$\nu = 1/n \sum \hat{P}_{i,s}(Y_{i,s} - \hat{P}_{i,s}).$$

In fact there holds a more general form of orthogonality

$$\hat{p}_s^T e_t/n \xrightarrow{p} 0 \quad \forall s, t, \tag{23}$$

and instead of (7) we have

$$\hat{P}^T E/n \xrightarrow{p} 0. \tag{24}$$

This is so because with appropriate changes in notation the argument of Section 6 applies line by line to

$$\nu^* = 1/n \sum \hat{P}_{i,s}(Y_{i,t} - \hat{P}_{i,t}).$$

In terms of $\hat{P}$ and Y this implies

$$\hat{P}^T Y \approx \hat{P}^T \hat{P}. \tag{25}$$

The matrix on the left occurs in sample enumeration where it gives the sums of predicted probabilities over outcome categories; it is here seen that it is (approximately) symmetrical. The sum of the estimated probabilities of alternative s over observations exhibiting outcome t is approximately equal to the sum of the probabilities of alternative t over observations exhibiting outcome s. This is not intuitively obvious.

References

Amemiya, T. (1985) *Advanced Econometrics.* Cambridge: Harvard University Press.

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) Assessing the fit of multinomial discrete models. *Unpublished paper.*

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

Windmeijer, Frank A.G. (1995) Goodness-of-fit measures in binary choice models. *Econometric Reviews* **14**, 101-116.