Categorical Data

A. Colin Cameron
University of California, Davis

March 02, 2006

Paper # 06-12

A very brief survey of regression for categorical data. Categorical outcome (or discrete outcome or qualitative response) regression models are models for a discrete dependent variable recording in which of two or more categories an outcome of interest lies. For binary data (two categories) probit and logit models or semiparametric methods are used. For multinomial data (more than two categories) that are unordered, common models are multinomial and conditional logit, nested logit, multinomial probit, and random parameters logit. The last two models are estimated using simulation or Bayesian methods. For ordered data, standard multinomial models are ordered logit and probit, or count models are used if ordered discrete data are actually a count.

Categorical Data

A. Colin Cameron
Department of Economics, University of California, Davis

March 2, 2006

Abstract

Categorical outcome (or discrete outcome or qualitative response) regression models are models for a discrete dependent variable recording in which of two or more categories an outcome of interest lies. For binary data (two categories) probit and logit models or semiparametric methods are used. For multinomial data (more than two categories) that are unordered, common models are multinomial and conditional logit, nested logit, multinomial probit, and random parameters logit. The last two models are estimated using simulation or Bayesian methods. For ordered data, standard multinomial models are ordered logit and probit, or count models are used if ordered discrete data are actually a count.

Keywords: Bayesian methods; binary data; bivariate probit; categorical data; choice-based sampling; conditional logit; count data; discrete outcome; extreme value distribution; index model; independence of irrelevant alternatives; latent variable; limited dependent variable; logit; log-odds ratio; logistic distribution; marginal effect; maximum score estimator; multinomial data; multinomial logit; multinomial probit; multivariate outcomes; nested logit; ordered logit; ordered probit; ordinary least squares; panel categorical data; Poisson regression; probit; qualitative response model; random parameters logit; random utility model; simulation-based estimation.

JEL Classification: C21, C25

Prepared for *New Palgave Dictionary of Economics*, 2nd edition.

1 Introduction

Categorical outcome models are regression models for dependent variable that is a discrete variable recording in which of two or more categories, usually mutually exclusive, an outcome of interest lies.

Categorical outcome models are also called discrete outcome models or qualitative response models, and are an example of a limited dependent variable model. Different models specify different functional forms for the probabilities of each category. These models are binomial or multinomial models, usually estimated by maximum likelihood.

Key early econometrics references include McFadden (1974), Amemiya (1981), Manski and McFadden (1981) and Maddala (1983). For textbook treatments see Amemiya (1985), Wooldridge (2002), Greene (2003) and Cameron and Trivedi (2005). The recent econometrics literature has focused on semiparametric estimation, see Pagan and Ullah (1999), and on simulation-based estimation of multinomial models, see Train (2003).

2 Binary Outcomes: Logit and Probit

Binary outcomes provide the simplest case of categorical data, with just two possible outcomes. An example is whether or not an individual is employed and whether or not a consumer makes a purchase.

For binary outcomes the dependent variable y takes one of two values, for simplicity coded as 0 or 1. If $y_i = 1$ with probability p_i, then necessarily $y_i = 0$ with probability $1 - p_i$, where i denotes the i^{th} of N observations. Regressors $\mathbf{x}_i$ are introduced by parameterizing the probability p_i, with

$$p_i = \Pr[y_i = 1|\mathbf{x}_i] = F(\mathbf{x}_i'\boldsymbol{\beta}),$$

where $F(\cdot)$ is a specified function and a single-index form is assumed.

The obvious choice of $F(\cdot)$ is a cumulative distribution function (cdf) since this ensures that $0 < p_i < 1$. The two standard models are the logit model with $p_i = \Lambda(\mathbf{x}_i'\boldsymbol{\beta}) = e^{\mathbf{x}_i'\boldsymbol{\beta}}/(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})$, where $\Lambda(z) = e^z/(1 + e^z)$ is the logistic cdf, and the probit model with $p_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$, where $\Phi(\cdot)$ is the standard normal cdf.

Interest usually lies in the marginal effect of a change in regressor on the probability that $y = 1$. For the r^{th} regressor, $\partial p_i/\partial x_{ir} = F'(\mathbf{x}_i'\boldsymbol{\beta})\beta_r$ where F' denotes the derivative of F. The sign of β_r gives the sign of the marginal

effect, if F is a continuous cdf since then $F' > 0$, though the magnitude depends on the point of evaluation $\mathbf{x}_i$. Common methods are to report the average marginal effect over all observations or to report the marginal effect evaluated at $\bar{\mathbf{x}}$.

Parameter estimates are usually obtained by maximum likelihood (ML) estimation. Given p_i, the density can be conveniently expressed as $f(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$. Assuming independence over i the resulting log-likelihood function is

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1-y_i)\ln(1-F(\mathbf{x}_i'\boldsymbol{\beta}))\}.$$

It can be shown that consistency of the MLE requires only that $p_i = F(\mathbf{x}_i'\boldsymbol{\beta})$, i.e. that the functional form for the conditional probability is correctly specified.

There is usually little difference between the predicted probabilities obtained by probit or logit, except for very low and high probability events. For the logit model $\ln[p_i/(1-p_i)] = \mathbf{x}_i'\boldsymbol{\beta}$, so that β_r gives the marginal effect of a change in x_{ir} on the log-odds ratio, a popular interpretation in the biostatistics literature.

A simpler method for binary data is OLS regression of y_i on $\mathbf{x}_i$, with White heteroskedastic robust standard errors used to control for the intrinsic heteroskedasticity in binary data. A serious defect is that OLS permits predicted probabilities to lie outside the $(0,1)$ interval. But it can be useful for exploratory analysis, as OLS coefficients can be directly interpreted as marginal effects and standard methods then exist for complications such as endogenous regressors.

When one of the outcomes is uncommon, surveys may oversample that outcome. For example, a survey of transit use may be taken at bus stops to oversample bus riders. This is a leading example of choice-based sampling. Standard ML estimators are inconsistent and instead one must use alternative estimators such as appropriately weighted ML.

The preceding discussion presumes knowledge of F. A considerable number of semiparametric estimators that provide consistent estimates of $\boldsymbol{\beta}$ given unknown F have been proposed. Manski's (1975) smooth maximum score estimator was a very early example of semiparametric estimation.

Index Models

Define a latent (or unobserved) variable y_i^* that measures the propensity for the event of interest to occur. If y_i^* crosses a threshold, normalized to be zero, then the event occurs and we observe $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$. If $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i$, then

$$p_i = \Pr[y_i^* > 0] = \Pr[-u_i < \mathbf{x}_i'\boldsymbol{\beta}] = F(\mathbf{x}_i'\boldsymbol{\beta}),$$

where $F(\cdot)$ is the cdf of $-u_i$.

The logit model arises if u_i has the logistic distribution. The probit model arises if u_i has the more obvious standard normal distribution, where imposing a unit error variance ensures model identification. The probit model ties in nicely with the Tobit model, where more data are available and we actually observe $y_i = y_i^*$ when $y_i^* > 0$. And it extends naturally to ordered multinomial data.

Random Utility Models

In many economics applications the binary outcome is determined by individual choice, such as whether or not to work. Then the outcome should be the alternative with highest utility. The additive random utility model (ARUM) specifies the utility for individual i of alternative j to be $U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}_j + \varepsilon_{ij}$, $j = 0, 1$, where the error term captures factors known by the decision-maker but not the econometrician. Then

$$p_i = \Pr[U_{i1} > U_{i0}] = \Pr[(\varepsilon_{i0} - \varepsilon_{i1}) \leq \mathbf{x}_{i1}'\boldsymbol{\beta}_1 - \mathbf{x}_{i0}{}'\boldsymbol{\beta}_0] = F(\mathbf{x}_{i1}'\boldsymbol{\beta}_1 - \mathbf{x}_{i0}{}'\boldsymbol{\beta}_0)$$

where F is the cdf of $(\varepsilon_{i0} - \varepsilon_{i1})$. For components x_{ir} of $\mathbf{x}_i$ that vary across alternatives (so $x_{i0r} \neq x_{i1r}$) it is common to restrict $\beta_{0r} = \beta_{1r} = \beta_r$. For components x_{ir} of $\mathbf{x}_i$ that are invariant across alternatives (so $x_{i0r} = x_{i1r}$) only the difference $\beta_{1r} - \beta_{0r}$ is identified.

The probit model arises, after rescaling, if ε_{i0} and ε_{i1} are iid standard normal. The logit model arises if ε_{i0} and ε_{i1} are iid type 1 extreme value distributed with density $f(\varepsilon) = e^{-\varepsilon}\exp(-e^{-\varepsilon})$. The latter less familiar distribution provides more tractable results when extended to multinomial models.

3 Multinomial Outcomes

Multinomial outcomes occur when there are more than two categorical outcomes. With m outcomes the dependent variable y takes one of m mutually exclusive values, for simplicity coded as $1, ..., m$. Let p_j denote the probability that the j^{th} outcome occurs. The multinomial density for y can be written as $f(y) = \prod_{j=1}^{m} p_j^{y_j}$ where y_j, $j = 1, ..., m$, are m indicator variables equal to 1 if $y = j$ and equal to 0 if $y \neq j$. Introducing a further subscript for the i^{th} individual and assuming independence over i yields log-likelihood

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij},$$

where the probabilities p_{ij} are modelled to depend on regressors and unknown parameters $\boldsymbol{\beta}$.

There are many different multinomial models, corresponding to different parameterizations of p_{ij}.

Unordered Multinomial Models

Usually the outcomes are unordered, such as in choice of transit mode to work. The benchmark model for unordered outcomes is the multinomial logit model. When regressors vary across alternatives (such as prices), the conditional logit (CL) model specifies $p_{ij} = e^{\mathbf{x}_{ij}'\boldsymbol{\beta}} / \sum_{k=1}^{m} e^{\mathbf{x}_{ik}'\boldsymbol{\beta}}$. If regressors are invariant across alternatives (such as gender), the multinomial logit (MNL) model specifies $p_{ij} = e^{\mathbf{x}_i'\boldsymbol{\beta}_j} / \sum_{k=1}^{m} e^{\mathbf{x}_i'\boldsymbol{\beta}_k}$, with a normalization such as $\boldsymbol{\beta}_1 = \mathbf{0}$ to ensure identification. In practice some regressors may be a mix of invariant and varying across alternatives; such cases can be re-expressed as either a CL or MNL model.

The CL and MNL models reduce to a series of pairwise choices that do not depend on the other choices available. For example, the choice between use of car or red bus is not affected by whether another alternative is a blue bus (essentially the same as the red bus). This restriction, called the assumption of independence of irrelevant alternatives, has led to a number of alternative models.

These models are based on the ARUM. Suppose the j^{th} alternative has utility $U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}$, $j = 1, ..., m$. Then

$$p_{ij} = \Pr[U_{ij} \geq U_{ik} \text{ for all } k] = \Pr[(\varepsilon_{ik} - \varepsilon_{ij}) \leq (\mathbf{x}_{ij}'\boldsymbol{\beta} - \mathbf{x}_{ik}'\boldsymbol{\beta}) \ \forall \ k].$$

The CL and MNL models arise if the errors ε_{ij} are iid type 1 extreme value distributed. More general models permit correlation across alternatives j in the errors ε_{ij}.

The most tractable model with error correlation is a nested logit model. This arises if the errors are generalized extreme value distributed. This model is simple to estimate but suffers from the need to specify a particular nesting structure.

The richer multinomial probit model specifies the errors to be $m-$dimensional multivariate normal with $(m+1)$ restrictions on the covariances to ensure identification. In practice it has proved difficult to jointly estimate both $\boldsymbol{\beta}$ and the covariance parameters in this model. A recent popular model is the random parameters logit model. This begins with a multinomial logit model but permits the parameters $\boldsymbol{\beta}$ to be normally distributed. For these two models there is no closed form expression for the probabilities and estimation is usually by simulation methods or Bayesian methods.

Ordered Multinomial Models

In some cases the outcomes can be ordered, such as health status being excellent, good, fair or poor.

The starting point is an index model, with single latent variable, $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i$. As y^* crosses a series of increasing unknown thresholds we move up the ordering of alternatives. For example, for $y^* > \alpha_1$ health status improves from poor to fair, for $y^* > \alpha_2$ it improves further to good, and so on. For the ordered logit (probit) model the error u is logistic (standard normal) distributed.

An alternative model is a sequential model. For example one may first decide whether or not to go to college ($y = 1$) and if chose college then choose either two-year college ($y = 2$) or four-year college ($y = 3$). The two decisions may be modelled as separate logit or probit models.

A special case of ordered categorical data is a count, such as number of visits to a doctor taking values 0, 1, 2, ... An ordered model can be applied to these data, but it is better to use count models. The simplest count model is Poisson regression with exponential conditional mean $\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. Common procedures are to use the Poisson but obtain standard errors that relax the Poisson restriction of variance-mean equality, to estimate the richer negative binomial model, or to estimate hurdle or two-part models or with-zeroes models that permit the process determining zero counts to differ from that for positive counts.

4 Multivariate Outcomes and Panel Data

Multivariate discrete data arise when more than one discrete outcome is modelled. The simplest example is bivariate binary outcome data. For example, we may seek to explain both employment status (work or not work) and family status (children or no children). The standard model is a bivariate probit model that specifies an index model for each dependent variable with normal errors that are correlated. Such models can be extended to permit simultaneity.

For panel binary data the standard model is an individual specific effects model with $p_{it} = F(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta})$ where α_i is an individual specific effect. The random effects model usually specifies $\alpha_i \sim N[0, \sigma^2_\alpha]$ and is estimated by numerically integrating out α_i using Gaussian quadrature. The fixed effects model treats α_i as a fixed parameter. In short panels with few time periods consistent estimation of $\boldsymbol{\beta}$ is possible in the fixed effects logit but not the fixed effects probit model. If $\mathbf{x}_{it}$ includes $y_{i,t-1}$, a dynamic model, fixed effects logit is again possible but requires four periods of data.

5 Bibliography

Amemiya, T. 1981. Qualitative Response Models: A Survey. Journal of Economic Literature, 19, 1483-1536.

Amemiya, T. 1985. Advanced Econometrics. Cambridge, MA: Harvard University Press.

Cameron. A.C. and P.K. Trivedi. 2005. Microeconometrics: Methods and Applications. Cambridge, UK: Cambridge University Press.

Greene, W.H. 2003. Econometric Analysis, fifth edition. Upper Saddle River, NJ: Prentice-Hall.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics, ed. P. Zarembka, New York: Academic Press, 105–142.

Maddala, G.S. 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge, UK: Cambridge University Press.

Manski, C.F. 1975. The maximum score estimator of the stochastic utility model of choice. Journal of Econometrics 3, 205–228.

Manski, C.F. and McFadden, D. (eds.) 1981. Structural Analysis of Discrete Data with Econometric Applications. Cambridge, MA.: MIT Press.

Pagan, A.R. and A. Ullah. 1999. Nonparametric Econometrics. Cambridge, UK: Cambridge University Press.

Train, K.E. 2003. Discrete Choice Methods with Simulation. Cambridge, UK: Cambridge University Press.

Wooldridge, J.M. 2002. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.