Estimation of a Binary Choice Model with Grouped Choice Data

Lyubov A. Kurkalova and Sergey Rabotyagov

Working Paper 05-WP 381
January 2005

Center for Agricultural and Rural Development
Iowa State University
Ames, Iowa 50011-1070
www.card.iastate.edu

Lyubov Kurkalova is an associate scientist and Sergey Rabotyagov is a graduate assistant in the Resource and Environmental Policy Division of the Center for Agricultural and Rural Development at Iowa State University.

The authors acknowledge helpful discussions with Wayne Fuller, Michael W. McCracken, Ivan Jeliazkov, Joseph Herriges, and Catherine Kling. Remaining errors are attributable to the authors.

For questions or comments about the contents of this paper, please contact Lyubov Kurkalova, 560A Heady Hall, Iowa State University, Ames, IA 50011-1070; Phone: 515-294-7695; Fax: 515-294-6336; E-mail: lyubov@iastate.edu.

Abstract

We propose an econometric technique for estimating the parameters of a binary choice model when only aggregated data are available on the choices made. The method performs favorably in applications to both simulated and real world choice data.

Keywords: aggregated choice data, binary choice model.

JEL Code: C25

ESTIMATION OF A BINARY CHOICE MODEL WITH GROUPED CHOICE DATA

Introduction

In many areas of economics, interest centers on the estimation of binary choice models and related issues, yet information on the choices made by individuals may be costly to collect or the choice data may be inaccessible to researchers because of confidentiality concerns. On the other hand, analysts may have access to the choice data aggregated across groups of individuals in the form of counts or proportions. If the observed predictors of the choice do not vary within the groups, individual choice models are easily estimable with such aggregated choice data (Greene 2004; Maddala 1983). However, when the predictors vary within the groups, the prevailing approach has been to abandon discrete choice models and use group averages of the predictors to estimate the models, explaining the *grouped* and not *individual* choices (e.g., Miller and Plantinga 1999). This study shows that neither avoiding individual choice models nor losing information by averaging over the individual-level predictors is necessary. We present an econometric method for estimation of a binary choice model when information on the attributes of the decisionmakers is available at the individual level, but the information on the choices made is aggregated across groups of individuals. The likelihood function that allows for this type of data is constructed and estimation of the resulting model using the method of maximum likelihood is proposed. We illustrate the method in a simulation study and in an application to a model of conservation tillage adoption.

Model and Method

Consider a set of N observations corresponding to binary choices made by N individuals. The choice is described by the variable Y_i, which takes on the value of 1 or 0 depending on whether a certain alternative A is adopted (chosen), that is,

$$Y_i = \begin{cases} 1, & \textit{if} \quad A \textit{ is adopted}, \\ 0, & \textit{otherwise}, \end{cases} \quad i = 1,...,N .$$

Choice is a function of K predictors, representing the attributes of the decisionmaker and/or the choices. The predictors are given by the vector $\mathbf{x}_i = (x_{1i},...,x_{Ki})$, $i = 1,...,N$. As in a standard econometrics setting, the exact relationship between Y_i and $\mathbf{x}_i$ is assumed known to the individuals making the choice but unobservable by researchers. As a consequence, the probability of adopting A from the researchers' perspective can be specified as $Pr[Y_i = 1] = Pr[\varepsilon_i < h(\boldsymbol{\beta}, \mathbf{x}_i)]$, where ε_i represents the researchers' error arising from measurement errors, functional form choice errors, and the effect of omitted variables; $h(.)$ is a specified function of its parameters, and vector $\boldsymbol{\beta}$ represents the parameters of interest. We assume ε_i to be independent across *i*, each ε_i to be logistically distributed, and $h(.)$ to have a linear functional form. Thus,

$$Pr[Y_i = 1] = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i)}. \tag{1}$$

As will become clear later, only the independence of ε_i is crucial for the proposed method; the other two assumptions can be easily relaxed. When the data y_i on the choices Y_i and the predictors $\mathbf{x}_i$ are available for all *i*, model (1) can be conventionally estimated using the method of maximum likelihood.

We now turn to the case in which less information on choices made is available to researchers. Specifically, we assume that $\mathbf{x}_i$ is still observed for all *i*, but instead of the observations on Y_i, only the sums (or averages) of the observations on Y_i over certain groups of individuals are available. That is, the observations $\bar{y}^{G_j}$ are available on the random variables $\bar{Y}^{G_j} \equiv \sum_{i \in G_j} Y_i$, where G_j are mutually exclusive, non-empty subsets of $\{1,...,N\}$ such that $\bigcup_j G_j = \{1,...,N\}$; N^{G_j} is the number of observations in group G_j,

j=1,…, J; and $\sum_{i \in G_j} N^{G_j} = N$. The quantities $\overline{Y}^{G_j} / N^{G_j}$ are usually interpreted as the proportions (shares) of the corresponding groups adopting A (e.g., Greene 2004).

Although important caveats exist (Garrett 2003), if the model we are considering were linear in the parameters of interest, this structure of the data would not create a serious problem for identification of the parameters. Indeed, for a linear counterpart of model (1) given by $Y_i = \mathbf{b}'\mathbf{x}_i + \eta_i$, where η_i are independently and identically distributed (i.i.d.) error terms, one can estimate the parameters $\mathbf{b}$ by fitting the model with aggregate data, that is, by fitting the model $\overline{Y}^{G_j} / N^{G_j} = \mathbf{b}'\overline{\mathbf{x}}^{G_j} + \overline{\eta}^{G_j}$, where $\overline{\mathbf{x}}^{G_j} \equiv \frac{1}{N^{G_j}} \sum_{i \in G_j} \mathbf{x}_i$ and $\overline{\eta}^{G_j} \equiv \frac{1}{N^{G_j}} \sum_{i \in G_j} \eta_i$. The inherent nonlinearity in model parameters precludes using a similar approach for the binary choice model (1).

In the literature (e.g., Miller and Plantinga 1999), the grouped choice data have been routinely paired with group average predictors data to estimate logistic models:

$$\overline{Y}^{G_j} / N^{G_j} = \frac{\exp\left(\boldsymbol{\alpha}'\overline{\mathbf{x}}^{G_j}\right)}{1 + \exp\left(\boldsymbol{\alpha}'\overline{\mathbf{x}}^{G_j}\right)} + \xi_j, \tag{2}$$

where ξ_j are i.i.d. error terms. While this model is useful for explaining and predicting *grouped* choices (the proportions of individuals adopting A), it is not immediately useful for explaining and predicting *individual* choices (which individuals adopt A). This is because of the nonlinearity of the postulated relationship: the parameters $\boldsymbol{\alpha}$ in equation (2) cannot be interpreted as parameters $\boldsymbol{\beta}$ in equation (1).

The method of recovery of parameters $\boldsymbol{\beta}$ we propose builds on the observation that, given the assumed independence of ε_i, the probability $\Pr\left[\overline{Y}^{G_j} = \overline{y}^{G_j}\right]$ can be represented as the sum of $\binom{N^{G_j}}{\overline{y}^{G_j}}$ number of terms, the probabilities of disjoint events in each of which exactly $\overline{y}^{G_j}$ of N^{G_j} individuals adopt A. Thus, the probability $\Pr\left[\overline{Y}^{G_j} = \overline{y}^{G_j}\right]$ can

be expressed in terms of the original parameters of interest $\boldsymbol{\beta}$ and data on $\mathbf{x}_i$. For example, for $N^{G_j} = 3$,

$$\begin{aligned}
&\Pr\left[\bar{Y}^{G_j} = 1\right] = \Pr\left[1 \textit{ out of } 3 \textit{ individuals in group } G_j \textit{ adopt } A\right] \\
&= \Pr\left[1^{st} \textit{ adopts } A,\ 2^{nd} \textit{ and } 3^{d} \textit{ do not adopt } A\right] + \Pr\left[2^{nd} \textit{ adopts } A,\ 1^{st} \textit{ and } 3^{d} \textit{ do not adopt } A\right] \\
&+ \Pr\left[3^{d} \textit{ adopts } A,\ 1^{st} \textit{ and } 2^{nd} \textit{ do not adopt } A\right] \\
&= \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_1)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_1)} \bullet \frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_2)} \bullet \frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_3)} + \frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_1)} \bullet \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_2)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_2)} \bullet \frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_3)} \\
&+ \frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_1)} \bullet \frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_2)} \bullet \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_3)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_3)}.
\end{aligned}$$

Following this line of thought, the likelihood function for the *j*th group of observations in general can be written as

$$L\left(\boldsymbol{\beta} \mid \bar{y}^{G_j}, N^{G_j}, \mathbf{x}_i\left(i \in G_j\right)\right) = \sum_{\sum_{i=1}^{N^{G_j}} \delta_i = \bar{y}^{G_j}} \prod_{i=1}^{N^{G_j}} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^{\delta_i} \left(\frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^{1-\delta_i}, \quad (3)$$

where δ_i takes on the value of 0 or 1 and plays the role of the unobserved information on the individual choices y_i.

Note that if predictors $\mathbf{x}_i$ do not vary within groups, that is, if $\mathbf{x}_i = \mathbf{x}^{G_j}$ for all $i \in G_j$, then the likelihood function (3) collapses to

$$L\left(\boldsymbol{\beta} \mid \bar{y}^{G_j}, N^{G_j}, \mathbf{x}_i\left(i \in G_j\right)\right) = \binom{N^{G_j}}{\bar{y}^{G_j}} \frac{\left(\exp\left(\beta' x^{G_j}\right)\right)^{\bar{y}^{G_j}}}{\left(1+\exp\left(\beta' x^{G_j}\right)\right)^{N^{G_j}}},$$

which is consistent with the log likelihood reported for this case by Greene (2004, p. 836).

We propose estimation of parameters $\boldsymbol{\beta}$ of model (1) by applying the method of maximum likelihood to the likelihood function (3). Next we present empirical applications of the approach described.

Simple Simulation Exercise

To demonstrate the proposed technique on simulated data, we set $K = 2$, $N = 10{,}000$, randomly draw "independent variables" $x_{1i}, x_{2i}, i = 1,...,10{,}000$, from $Unif(0,20)$ distribution, and first let $(\beta_1, \beta_2) = (-1,1)$. Then, we randomly draw $u_i, i = 1,...,10{,}000$ from $Unif(0,1)$ distribution and obtain the sample of $y_i, i = 1,...,10{,}000,$ using the inverse cumulative density function (cdf) method:

$$y_i = \begin{cases} 1, & if \quad u_i \le \dfrac{\exp(\beta_1 x_{1i} + \beta_2 x_{2i})}{1 + \exp(\beta_1 x_{1i} + \beta_2 x_{2i})}, \\ 0, & otherwise. \end{cases}$$

We additionally consider three more pairs of (β_1, β_2), as reported in Table 1. The choice of the parameter vector values, although arbitrary in principle, in this case was made so that the average probability of adoption A varies from 5 to 50 percent. The grouped choice data are constructed by randomly grouping the observations into 5,000, 2,000, or 1,000 groups and summing the y_i's over the groups. The results of model (1) estimation under four alternative assumptions on availability of the choice data are reported in Table 1.

We find that the parameter estimates obtained from the grouped data settings are close both to the true parameters and to those estimated when individual choices are observed. We also find that, for any true parameter, as the number of groups decreases (i.e., the number of individuals per group increases), the estimated standard errors increase. This finding is intuitively appealing: the more aggregated the grouped data are, the less information is available to recover the parameter values, thus increasing the estimation uncertainty represented by the standard errors. Not surprisingly, when the data are divided into 5,000 groups, the proposed technique performs best across various adoption probabilities. In this case, the data structure most closely resembles that of a standard binary choice model, as on average there are only two observations per group. A similar decline in estimation precision as one moves from individual to grouped choice data has been demonstrated by Warner (as reported in Maddala 1983, p. 32), although the only grouping considered in that study is the one under which the predictors did not vary within the groups.

TABLE 1. Estimation on simulated data, 10,000 observations (standard errors in parentheses)

Average Adoption Rate	True Parameters		Estimates, Individual Choices Observed	Estimates, Only Grouped Choices Observed		
				5,000 Groups	2,000 Groups	1,000 Groups
5%	β_1	-1	-0.984 (0.039)	-0.992 (0.047)	-0.984 (0.049)	-0.994 (0.060)
	β_2	0.052	0.0501 (0.0067)	0.0511 (0.0073)	0.0488 (0.0078)	0.0492 (0.0093)
10%	β_1	-1	-0.984 (0.031)	-1.003 (0.035)	-0.999 (0.038)	-0.988 (0.046)
	β_2	0.18	0.1744 (0.0075)	0.1783 (0.0080)	0.1767 (0.0088)	0.173 (0.010)
30%	β_1	-1	-1.033 (0.028)	-1.062 (0.032)	-1.066 (0.035)	-1.065 (0.039)
	β_2	0.59	0.610 (0.017)	0.629 (0.019)	0.634 (0.021)	0.633 (0.024)
50%	β_1	-1	-1.035 (0.028)	-1.031 (0.030)	-1.036 (0.033)	-1.021 (0.039)
	β_2	1	1.040 (0.028)	1.035 (0.030)	1.042 (0.033)	1.028 (0.039)

Application to a Model of Conservation Tillage Adoption

We apply the proposed technique to estimation of a model of conservation tillage adoption similar to that of Kurkalova, Kling, and Zhao (2003). The model is derived under the assumption that a farmer will adopt conservation tillage if the expected annual net returns from it, π_1, exceed those from the alternative, conventional tillage, π_0, plus a premium, P, associated with uncertainty. Then, assuming that $\pi_1 - P$ is a linear function of a set of observed predictors $\mathbf{x}$ and that the observations on π_0 are available, the model takes the form

$$\Pr[adopt] = \Pr[\pi_1 \geq \pi_0 + P + \sigma\varepsilon] = \Pr\left[\varepsilon \leq \frac{\boldsymbol{\beta}'\mathbf{x}}{\sigma} - \frac{\pi_0}{\sigma}\right], \quad (4)$$

where ε is a logistic error. The parameters of interest are the linear function parameters $\boldsymbol{\beta}$, together with σ, the error term multiplier. The model is very useful for the modeling of adoption policy since the identification of both $\boldsymbol{\beta}$ and σ allows evaluation of the

opportunity cost of adoption for current non-adopters as well as prediction of the responsiveness of the probability of conservation tillage adoption to financial incentives (subsidies). Kurkalova, Kling, and Zhao (2003) estimate the model on data coming primarily from the 1992 National Resources Inventory (NRI) (Nusser and Goebel 1997) for the state of Iowa. However, traditional estimation of a similar binary choice model on 1997 (the latest available) NRI data is not possible, as the response variable, the indicator of adoption of conservation tillage, is not available in the 1997 NRI.

To estimate the conservation tillage adoption model on combined 1992 and 1997 Iowa NRI data (1,339 observations for 1992; 1,365 for 1997) we begin by grouping the observations by crop and county, which results in 240 groups for 1992 and 261 for 1997. While the 1992 group counts of adopters are obtainable from NRI, for each 1997 group, the counts of adopters in the group are constructed by rounding to the nearest integer the product of the number of observations in the group and the proportion of land in conservation tillage for the corresponding crop and county, which were taken from the Conservation Technology Information Center (http://www.ctic.purdue.edu/CTIC/CTIC.html) and Agricultural Resource Management Survey (http://www.ers.usda.gov/Briefing/ARMS) data.

The model estimated is given by

$$\Pr[Y_i] = \frac{\exp\left(\left(\sum_{j=1}^{22}\beta_j x_{ji} - \pi_{0,i}\right) / \left(\sigma^{92}\cdot I_{92,i} + \sigma^{97}\cdot\left(1 - I_{92,i}\right)\right)\right)}{1 + \exp\left(\left(\sum_{j=1}^{22}\beta_j x_{ji} - \pi_{0,i}\right) / \left(\sigma^{92}\cdot I_{92,i} + \sigma^{97}\cdot\left(1 - I_{92,i}\right)\right)\right)}, \tag{5}$$

with the predictors constructed as in Kurkalova, Kling, and Zhao 2003, to which we refer for the data interpretation. In addition to the net returns to conventional tillage ($\pi_{0,i}$), the predictors include land slope (x_{1i}), soil permeability (x_{2i}), average water-holding capacity of the soil (x_{3i}), means of daily maximum and minimum temperatures during the corn growing season (x_{4i} and x_{5i}, respectively), mean daily precipitation during corn growing season (x_{6i}), year 1992 indicator (1 if the year is 1992 and 0 otherwise) (x_{7i}), as well as 15 interaction terms defined as follows. Predictor x_{8i} is the product of the indicator for

corn (1 if the crop grown is corn and 0 otherwise) and the standard deviation of daily precipitation during corn growing season. Predictors x_{9i} - $x_{12,i}$ are the products of x_{8i} and net returns to conventional tillage, proportion of county operators working off-farm, county average farm operator age, and proportion of county operators that are male, respectively. Predictors $x_{13,i}$ - $x_{17,i}$ are constructed similarly to x_{8i} - $x_{12,i}$ except that instead of the indicator for corn, the indicator for soybeans is used. Finally, predictors $x_{18,i}$ - $x_{22,i}$ are also similar to x_{8i} - $x_{12,i}$, with the indicator for corn replaced with the indicator for crops other than corn or soybeans. Since choice data sources differ significantly between 1992 and 1997, we allowed the error term multiplier, parameter σ in equation (4), to vary by years; thus, $\sigma = \sigma^{92} \cdot I_{92,i} + \sigma^{97} \cdot \left(1 - I_{92,i}\right)$, where $I_{92,i}$ is the year 1992 indicator. The parameters of interest are the β's, together with σ^{92} and σ^{97}.

Model estimation results are provided in Table 2. Not surprisingly, the standard errors for σ^{97} are much larger than are those for σ^{92}, a finding that may reflect more noise in 1997 choice data, which comes not from direct summation of individual choice data but from a separate source that may be subject to an additional sampling error. We estimated the average (among current non-adopters) subsidy needed to induce adoption to be $12.36 in 1992 and $36.52 in 1997.

Conclusions

In this paper, we propose an econometric technique for recovering the parameters describing individual choices when only grouped data are available on the choices made. The method generalizes the grouped data models considered in the literature to the case when the predictors of the choices vary within the groups over which the aggregated choice data are reported. The model performed well in an application to simulated and real-world data. Importantly, it allowed us to obtain estimates relevant to policy analysis that incorporated the most recent data available, even though the structure of the data did not permit the application of conventional discrete-choice methods.

TABLE 2. Estimation results for the model of conservation tillage adoption

Parameter	Estimate	Standard Error	P-value
β_1	1.11	0.30	0.000
β_2	1.29	0.64	0.043
β_3	1.40	0.53	0.008
β_4	3.66	0.62	0.000
β_5	-4.18	0.67	0.000
β_6	143	80	0.073
β_7	76	22	0.001
β_8	2234	370	0.000
β_9	-2.66	0.08	0.000
β_{10}	-232	48	0.000
β_{11}	-7.7	1.9	0.000
β_{12}	-1290	290	0.000
β_{13}	1947	452	0.000
β_{14}	-3.30	0.20	0.000
β_{15}	-264	76	0.001
β_{16}	-7.8	2.2	0.001
β_{17}	-888	378	0.019
β_{18}	3107	1064	0.004
β_{19}	-3.64	0.56	0.000
β_{20}	-225	177	0.205
β_{21}	-17.3	9.4	0.067
β_{22}	-1424	1075	0.185
σ^{92}	10.3	1.6	0.000
σ^{97}	251	108	0.020

Note: Mean log-likelihood is -1.573.

References

Garrett, T.A. 2003. "Aggregated Versus Disaggregated Data in Regression Analysis: Implications for Inference." *Economics Letters* 81: 61-65.

Greene, W. 2004. *Econometric Analysis*, 4th ed. Upper Saddle River, NJ: Prentice Hall.

Kurkalova, L.A., C.L. Kling, and J. Zhao. 2003. "Green Subsidies in Agriculture: Estimating the Adoption Costs of Conservation Tillage from Observed Behavior." CARD Working Paper 01-WP 286. Center or Agricultural and Rural Development, Iowa State University. March.

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

Miller, D.J., and A.J. Plantinga. 1999. "Modeling Land Use Decisions with Aggregate Data." *American Journal of Agricultural Economics* 81(2): 180-94.

Nusser, S.M., and J.J. Goebel. 1997. "The National Resources Inventory: A Long-Term Multi-Resource Monitoring Programme." *Environmental and Ecological Statistics* 4: 181-204.