A MAXIMUM LIKELIHOOD METHOD FOR LATENT CLASS REGRESSION INVOLVING A CENSORED DEPENDENT VARIABLE

KAMEL JEDIDI
MARKETING DEPARTMENT
GRADUATE SCHOOL OF BUSINESS
COLUMBIA UNIVERSITY

VENKATRAM RAMASWAMY
MARKETING DEPARTMENT
SCHOOL OF BUSINESS ADMINISTRATION
UNIVERSITY OF MICHIGAN

WAYNE S. DESARBO
MARKETING AND STATISTICS DEPARTMENTS
SCHOOL OF BUSINESS ADMINISTRATION
UNIVERSITY OF MICHIGAN

The standard tobit or censored regression model is typically utilized for regression analysis when the dependent variable is censored. This model is generalized by developing a conditional mixture, maximum likelihood method for latent class censored regression. The proposed method simultaneously estimates separate regression functions and subject membership in K latent classes or groups given a censored dependent variable for a cross-section of subjects. Maximum likelihood estimates are obtained using an EM algorithm. The proposed method is illustrated via a consumer psychology application.

Key words: censored regression, latent class analysis, maximum likelihood estimation, consumer psychology.

1. Introduction

Tobit models are part of a general class of models for analyzing truncated and censored data where the range of the dependent variable is constrained (see Amemiya, 1984, for a survey). A typical case of censoring occurs when the dependent variable has a number of its values concentrated at a limiting value, say zero. For instance, in a large-scale study of the number of hours worked by married women (Greene & Quester, 1982), about 66% of the over 10,000 wives surveyed reported zero hours. This makes the use of the classical regression model inappropriate, as to be discussed below.

The standard tobit or censored regression model (Tobin, 1958) in which the dependent variable is censored (at zero, without loss of generality) can be expressed as

\[ y_i^* = x_i' \beta + u_i, \quad i = 1, \ldots, I \]  

where the random variable \( y_i^* \) may be viewed as an index or a partially latent variable whose observed value, \( y_i \), is concentrated at zero when it is nonpositive. Hence,
$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0, \\ 0 & \text{if } y_i^* \leq 0, \end{cases}$

where $y_i$ is the value of the observed censored dependent variable for subject $i$. In (1), $x_i^j$ is the $i$-th row vector of $x = (x_{ij})$ containing the values of $J$ explanatory variables ($j = 1, \ldots, J$) for subject $i$. The regression coefficients for these explanatory variables are contained in $\beta = \{\beta_j\}$, a $J \times 1$ column vector, and the error terms $\{u_i\}$ are assumed to be iid drawings from a normal distribution, $N(0, \sigma^2)$. Note that $y_i$ and $x_i$ are known for each of the $I$ subjects ($i = 1, \ldots, I$), but $y_i^*$ is unobserved if it is nonpositive (i.e., $y_i = 0$) and is therefore partially latent.

The expected value of $y_i = E(y_i) = \text{Prob } (y_i^* > 0) \cdot E(y_i|y_i^* > 0)$, and the conditional expectation is given by (see Amemiya, 1985, p. 367)

$$E(y_i|y_i^* > 0) = x_i^\prime \beta + E(u_i|u_i > -x_i^\prime \beta)$$

$$= x_i^\prime \beta + \sigma \left( \phi \left( \frac{x_i^\prime \beta}{\sigma} \right) \right), \quad (2)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution functions, respectively, of a standard normal variable. Note that (2) implies that the classical regression estimator $\hat{\beta}$ is biased and inconsistent regardless of whether one uses all the observations or just the positive observations. Thus, the classical regression model is inappropriate when the dependent variable is censored.

Numerous applications of standard tobit models for censored data have appeared in the social science literature, in part due to the increasing availability of computational resources as well as micro-level survey and panel data. These include, inter alia, psychological, economic, and social research studies concerning family relations and attachment behavior (Fisher & Tesler, 1986), number of extramarital affairs (Fair, 1978), job absenteeism (Baba, 1990), personal wealth transfers (Adams, 1980), ratio of unemployed hours to employed hours (Ashenfelter & Ham, 1979), household purchases of durable goods (Tobin, 1958), number of credit card accounts possessed by consumers (Kinsey, 1981), number of criminal arrests (Witte, 1980), performance on achievement tests (Gross, 1980), and household purchases of grocery products (Elrod & Winer, 1982; Tellis, 1988).

In most of these applications, a single set of coefficients $\beta$ is estimated from the censored data. While this may be justified if one is only interested in aggregate-level estimates, it may be inadequate and potentially misleading if there is considerable heterogeneity in subjects' responses. Consider the study of the number of hours worked by married women (Greene & Quester, 1982) that investigates the impact of explanatory variables such as second marriage, divorce probability, presence of small kids, as well as education and wage differences between husband and wife. Aspects of these tobit analyses are reproduced in Table 1 (see Greene, 1990, pp. 728–729).

Separate tobit estimates are reported for black versus white wives. The coefficients in Table 1 suggest differences in the impact of the explanatory variables on the average number of hours worked. For instance, second marriage has a much larger positive impact for white wives than black wives, while education difference has a positive effect for black wives and a negative effect for white wives. Hence, if differences in responses exist across subjects, disaggregate analyses are necessary to reveal such differential effects. However, a priori bases for performing such disaggregate analyses...