Discrete Choice and Censoring

Whitney K. Newey

MIT

November, 2004

Discrete Choice

Multinomial Choice: Data consists of consumer choices of various goods, along characteristics. Let there be J choices and $y = (y_1, \ldots, y_J)$ where $y_j = 1$ if good j is chosen and $y_j = 0$ otherwise. Let x be observed characteristics of the goods and choices. Here a conditional density for y corresponds to conditional choice probabilities $P(j|x,\beta)$, one for each j, with $\sum_{j=1}^{J} P(j|x,\beta) = 1$ for all β and x. Then

$$\ln f(y|x,\beta) = \sum_{j=1}^{J} y_j \ln P(j|x,\beta).$$

For example, the multinmial logit model has $x = (x_1, \ldots, x_J)$ and

$$P(j|x,\beta) = \frac{e^{x_j'\beta}}{\sum_{k=1}^{J} e^{x_k'\beta}}.$$

This model has a random utility interpretation. If the utility of choice j is $x_j'\beta + \varepsilon_j$ where ε_j are i.i.d. over j with Type I Extreme Value distributions (with CDF $e^{-e^{-\varepsilon}}$), then the probability that j has the highest utility, and is thus chosen, has the form given above. McFadden used this to predict the effect of the introduction of BART on ridership of public and private transportation in the San Francisco Bay area.

One problem with this model is that $P(j|x,\beta)/P(k|x,\beta) = e^{x_j'\beta - x_k'\beta}$ depends only on the characteristics of alternatives j and k (this is called the independence from irrelevant alternatives property, or IIA). Approaches to deal with this include allowing β to be random and allowing ε_j to be correlated with each other. Allowing β to be random would lead to choice probabilities of the form

$$P(j|x,\beta) = \int \frac{e^{x_j'\gamma}}{\sum_{k=1}^{J} e^{x_k'\gamma}} h(\gamma|\beta)d\gamma.$$

Hard to compute. Multivariate normal ε_j probabilities (called multinomial probit) also hard to compute. A case with correlated ε_j that can be computed is nested logit. For $y = \ln(\exp(x_1'\beta/\lambda) + \exp(x_2'\beta/\lambda))$,

$$\begin{aligned} P(j|x,\beta) &= \frac{e^{x_j'\beta/\lambda} e^{\lambda y}}{e^{y}\left(e^{x_3'\beta} + e^{\lambda y}\right)}. \; j = 1, 2, \\ P(3|x,\beta) &= \frac{e^{x_3'\beta}}{e^{x_3'\beta} + e^{\lambda y}}. \end{aligned}$$

Multinomial logit on branches.

Duration Models

T: Lifetime or duration (e.g., unemployment, firm lifetimes).

x: Regressors (covariates).

Goal: Estimate effect of x on T; also estimate how conditional density of T depends on T.

Important general issue is censoring.

General parametric model: Let θ be a parameter vector, x regressors, and conditional survivor function

$$S(t \mid x, \theta) = \Pr(T \geq t \mid x, \theta)$$

Complete model for conditional distribution of T given x; other ways to describe this model.

$$\begin{array}{lcll} f(t \mid x, \theta) & = & -\frac{d}{dt} S(t \mid x, \theta); & \text{conditional pdf} \\ \lambda(t \mid x, \theta) & = & \frac{f(t|x,\theta)}{S(t|x,\theta)} = -\frac{d}{dt} \ln S(t \mid x, \theta); & \text{hazard rate} \\ \Lambda(t \mid x, \theta) & = & \int_0^t \lambda(t \mid x, \theta); & \text{integrated hazard} \end{array}$$

Relationships: above and $S(t|x,\theta) = \exp(-\Lambda(t|x,\theta))$.

We use the representation that is most convenient for a particular application. Some theories imply things about hazard, e.g. declining reservation wage in search theory implies $\frac{\partial \lambda}{\partial t}(t|x,\theta) > 0$, when T is the length of an umemployment spell.

Historically important class of models are proportional hazards

$$\lambda(t \mid x, \theta) = \lambda(t, \alpha) \exp(x'\beta), \quad \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Here changes in x just shift hazard up and down, i.e., shape of hazard as function of t entirely determined by $\lambda(t; \alpha)$

$$\frac{\lambda(t \mid x, \theta)}{\lambda(\tilde{t} \mid x, \theta)} = \frac{\lambda(t, \alpha)}{\lambda(\tilde{t}, \alpha)}.$$

Motivation: Convenient starting point, computationally, historically. Also implied by some theoretical models. See Heckman chapter, *Handbook of Econometrics, Volume 2*. $\lambda(t, \alpha)$ is called "baseline hazard".

Examples:

$$\begin{array}{ll} \lambda(t, \alpha) = \alpha; & \text{constant} \\ \lambda(t, \alpha) = \frac{\alpha_1 t^{\alpha_2 - 1}}{\alpha_2}; & \text{Weibull; allows } \frac{\partial \lambda}{\partial t} > 0, \frac{\partial \lambda}{\partial t} < 0 \end{array}$$

Censoring

Have to account for effects of sampling in likelihood. Longer spells will be more likely to appear in data when sampling at fixed points in time. See diagram. Length biased sampling. A general principle illustrated here is that ignoring sampling based on the endogenous variable will lead to inconsistent estimates.

Case 1. Random sample of completed spells.

Observations are

$$(T_1, x_1), \ldots, (T_n, x_n)$$

The MLE maximizes

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_i \ln f(T_i|x_i, \theta)$$

Almost never have data like this.

Case 2. Sample of unemployed people. Sample unemployed, ask them how long they have been unemployed, and then follow them until they are employed again. So, the likelihood must condition on the fact that they are unemployed when surveyed. If we know that they have been unemployed for t_i periods, then condition on $T_i \geq t_i$. (This setting analogous to that where only observe data when Y_i is positive). The MLE maximizes

$$\begin{aligned} \hat{Q}_n(\theta) &= \frac{1}{n} \sum_i \ln \left[\frac{f(T_i|x_i, \theta)}{S(t_i|x_i, \theta)} \right] \\ &= \frac{1}{n} \sum_i [\ln f(T_i|x_i, \theta) - \ln S(t_i|x_i, \theta)] \end{aligned}$$

Case 3. Right censoring. Do not observe completed spells for everyone. For some we just know duration is greater than c_i. For example, we survey unemployed once and find t_i, then survey them sometime later, and record when spell ended or whether they are still unemployed. Let $d_i = 1$ if complete spell is observed, $d_i = 0$ if only know lasts at least to c_i. The MLE maximizes

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_i [d_i \ln f(T_i|x_i, \theta) + (1 - d_i) S(c_i|x_i, \theta) - \ln S(t_i|x_i, \theta)]$$

Case 4. Discrete data. Do not know any durations. Only know whether spell is shorter or longer than c_i. Like when survey the second time we only know whether unemployment has ended.

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_i \{d_i \ln[S(t_i|x_i, \theta) - S(c_i|x_i, \theta)] + (1 - d_i) \ln S(c_i|x_i, \theta) - \ln S(t_i|x_i, \theta)\}$$

Discrete data like this is common in applications, e.g. weeks of unemployment data. We will discuss more later. Difficult to also include employed people at start; have to assume something about arrival rate.

Example: Lancaster (1979). British unemployment data. Consider effect of ignoring observed heterogeneity. Lancaster, Weibull proportional hazards; $\lambda(\hat{t};\alpha) = ct^{\hat{\alpha}_2 - 1}$. Lancaster Table III.

$\hat{\alpha}_2$	Regressors
.67	none
.74	ln(age);
.768	ln(age), ln(regional unemp), benefits.
.773	" " ", Earnings

Lancaster considered proportional hazards with heterogeneity. Let a proportional hazard model conditional on unobserved heterogeneity

$$S(t|x,\theta,v) = \exp(-\Lambda(t;\alpha)\exp(x'\beta)v)$$

Suppose we model survivor function as

$$S(t|x,\theta,\gamma) = \int S(t|x,\theta,v)f(v;\gamma)dv = M_v(-\Lambda(t,\alpha)\exp(x'\beta),\gamma),$$

where $M_v(s,\gamma)$ is the moment generating function of v. For instance, for a gamma density $f(v|\gamma) = v^{\gamma-1}e^{-v}/\Gamma(\gamma)$, the moment generating function is $M_v(s,\gamma) = (1-s)^{-\gamma}$, so that

$$S(t|x,\theta,\gamma) = [1 + \Lambda(t,\alpha)\exp(x'\beta)]^{-\gamma}.$$

When Lancaster estimates using this as the model for survivor function, he obtains $\hat{\alpha} = .90$, not significantly different than 1.

Ordered Data

Often duration data is discrete and ordered, like weeks of unemployment. Good model here is transformed regression model with unknown transformation (takes care of scaling problems). Also applies to duration data. This model specifies that there is an unknown, strictly increasing function $\tau(t)$ with

$$\tau(T) = -x'\beta + u,\ u \text{ has CDF } G(u,\gamma).$$

Here the function $\tau(t)$ is nonparametric, not having specified functional form. Thus, the model is semiparametric, in that it has a parametric and nonparametric part.

For the moment suppose we have completed spells. Can construct simple estimator. Subdivide $[0,\infty)$ into intervals $I_j = [t_{j-1}, t_j)$, $(j = 1,\ldots,J)$, with $t_0 = 0$, $t_J = \infty$. Let

$$y_j = \begin{cases} 1, & T \in I_j, \\ 0 & \text{otherwise} \end{cases}, (j = 1, \ldots, J),$$

and $\tau_j = \tau(t_j)$, $(j = 1, \ldots, J-1)$, $\tau_0 = -\infty$, $\tau_J = +\infty$. By $\tau(t)$ strictly increasing,

$$\begin{aligned} \Pr(y_j = 1|x) &= \Pr(t \in I_j|x) = \Pr(t_{j-1} \leq T < t_j|x) = \Pr(\tau_{j-1} \leq \tau(T) < \tau_j|x) \\ &= \Pr(\tau_{j-1} \leq -x'\beta + u < \tau_j) = \Pr(\tau_{j-1} + x'\beta \leq u < \tau_j + x'\beta) \\ &= G(\tau_j + x'\beta, \gamma_0) - G(\tau_{j-1} + x'\beta, \gamma_0) \stackrel{def}{=} P_j(x, \theta), \end{aligned}$$

where $\theta = (\beta', \gamma', \tau_1, \ldots, \tau_{J-1})'$. Note that these probabilities depend on just the parameters $\tau_1, \ldots, \tau_{J-1}$ and not the whole function. Thus, the likelihood is parametric, although the model is semiparametric. The log-likelihood of a single observation is

$$\ln f(y|x, \theta) = \sum_{j=1}^{J} y_j \ln P_j(x, \theta).$$

This is concave in β and $\tau_1, \ldots, \tau_n$ if the log of the density of $G(u, \gamma)$ is concave.

Proportional hazards specification is when $G(u, \gamma)$ is a mixture of Type I extreme value with some other distribution (i.e., $u = \varepsilon + \eta$, where ε is Type I extreme value; if not, then this is not proportional hazards model). In this case $\tau(t) = \ln \Lambda(t)$, so that $\tau_j = \ln \Lambda(t_j)$ and so a finite difference approximation to the hazard rate is

$$\hat{\lambda}(t_j) = (e^{\hat{\tau}_j} - e^{\hat{\tau}_{j-1}})/(t_j - t_{j-1}).$$

Censoring can be handled like before. To handle left censoring, we consider the conditional likelihood given that $T \geq t_{j_\ell}$, where t_{j_ℓ} is greater than or equal to the censoring point. For right censoring, we can consider the likelihood when we only know that one of $y_j = 1$ occurs for $T \geq t_{j_r}$. For $y_c = 1$ when censoring occurs and zero otherwise, the resulting log-likelihood is

$$\ln f(y|x, \theta) = (1 - y_c) \sum_{j_r > j \geq j_\ell} y_j \ln P_j(x, \theta) + y_c \ln[\sum_{j \geq j_r} P_j(x, \theta)] - \ln[\sum_{j \geq j_\ell}^{J} P_j(x, \theta)].$$

Han and Hausman (1990) give application. Data from PSID set created by Katz (1986). Waves 14 and 15. Interviewer asks whether unemployed last year and duration in weeks. Answer either length, or still unemployed. Thus no left censoring but still have right censoring. Actually have recalls or new jobs, treat same. Could treat different, see paper. For results, see tables and graphs.