1 Latent Variables Continued

Last time, we looked at the probit and logit models from two perspectives: a Bernoulli likelihood with \( p_i = F(x_i' \beta) \), and a latent variables approach. With the latter, we conceived of a regression of the form \( y_i^* = x_i' \beta + \varepsilon_i \), where \( y_i^* \) is a continuous variable we can’t observe completely. We only observe whether it’s positive or negative – which is reflected in the dummy variable \( y_i \). It turns out that if \( \varepsilon \sim N(0, \sigma^2) \), the probit model recovers estimates of \( \beta/\sigma \) from the latent model that converge asymptotically to their true values.

Probit and logit, it turns out, are just two examples of a broader class of models in which we assume a continuous, imperfectly observed measure.

2 Ordered Choice

Sometimes, we have ordinal data: discrete outcomes that can be ordered along a single dimension.

Examples:

- Survey response with Likert Scaling
- Electoral choice with three candidates who can be easily ranked along a single ideological dimension.

Consider a three choice survey model of ideology. There is an unobservable, continuous latent variable representing ideology, \( y_i^* \). Ideology is determined by a vector of exogenous variables and a \( N(0,1) \) error term \( \varepsilon \):

\[
y_i^* = x_i' \beta + \varepsilon_i
\]

(We can rescale the error to unit variance because there is no natural metric for ideology.)

A survey research asks respondents whether they consider themselves liberals, conservatives, or moderates. She records:
ideo = \begin{cases} 
1 & \text{if liberal} \\
2 & \text{if moderate} \\
3 & \text{if conservative}
\end{cases}

The research believes that respondents have categorized themselves as follows:

\[ y_i = \begin{cases} 
1 & \text{if } y_i^* \leq 0 \\
2 & \text{if } 0 < y_i^* \leq \tau \\
3 & \text{if } y_i^* > \tau
\end{cases} \]

where \( \tau \) is a “cutpoint” – a parameter to be estimated. Or:

\[ y_i = 1 \quad \text{if} \quad x_i' \beta + \varepsilon_i \leq 0 \]
\[ \varepsilon_i \leq -x_i' \beta \]
\[ y_i = 2 \quad \text{if} \quad 0 < x_i' \beta + \varepsilon_i \leq \tau \]
\[ -x_i' \beta < \varepsilon_i \leq \tau - x_i' \beta \]
\[ y_i = 3 \quad \text{if} \quad x_i' \beta + \varepsilon_i > \tau \]
\[ \varepsilon_i > \tau - x_i' \beta \]

How do we construct a ML estimator of

\[ \theta = \begin{bmatrix} \beta \\ \cdot \cdot \cdot \\ \tau \end{bmatrix} \]

Suppose \( y_i = 1 \), so \( \varepsilon_i \leq -x_i' \beta \). Then

\[ P(\varepsilon_i \leq -x_i' \beta) = \Phi(-x_i' \beta). \]

Now suppose \( y_i = 3 \), so \( \varepsilon_i > \tau - x_i' \beta \). Given the symmetry of \( \Phi(\cdot) \), we know:

\[ P(\varepsilon_i > \tau - x_i' \beta) = 1 - \Phi(\tau - x_i' \beta) = \Phi(x_i' \beta - \tau). \]

What about when \( y_i = 2 \). This occurs when \( -x_i' \beta < \varepsilon_i \leq \tau - x_i' \beta \), with probability

\[ P(-x_i' \beta < \varepsilon_i \leq \tau - x_i' \beta) = \Phi(\tau - x_i' \beta) - \Phi(-x_i' \beta). \]

We now have probabilities corresponding to each outcome, which enables us to construct a likelihood function. Let

\[ \delta_{i1} = 1_{y_i = 1} \]
\[ \delta_{i2} = 1_{y_i = 2} \]
\[ \delta_{i3} = 1_{y_i = 3} \]

\[ \{ \delta_{ij} \} \quad \text{where} \quad \sum_{j=1}^{3} \delta_{ij} = 1 \quad \forall i. \]
Then the likelihood for the full sample is

$$L(y|\beta, \tau) = \prod_{i=1}^{N} [\Phi(-x_i'\beta)]^{\delta_{i1}} \times [\Phi(\tau - x_i'\beta) - \Phi(-x_i'\beta)]^{\delta_{i2}} \times [\Phi(x_i'\beta - \tau)]^{\delta_{i3}}.$$  

The corresponding log-likelihood is

$$\ell(y|\beta, \tau) = \sum_{i=1}^{N} \{\delta_{i1} \ln [\Phi(-x_i'\beta)] + \delta_{i2} \ln [\Phi(\tau - x_i'\beta) - \Phi(-x_i'\beta)] + \delta_{i3} \ln [\Phi(x_i'\beta - \tau)]\}.$$  

This is called the ordered probit model.

Note (again) that because our data are ordinal, we can’t recover an estimate of $\sigma^2$.

We can add categories:

$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq 0 \\ 2 & \text{if } 0 < y_i^* \leq \tau_1 \\ 3 & \text{if } \tau_1 < y_i^* \leq \tau_2 \\ 4 & \text{if } y_i^* > \tau_2 \end{cases}$$  

We can estimate a model with no constant and an additional cutpoint, e.g.

$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < y_i^* \leq \tau_2 \\ 3 & \text{if } y_i^* > \tau_2 \end{cases}$$  

This no-constant approach will produce identical $\beta$s to the one described above. Stata uses the no-constant, $J$ cutpoint approach. Limdep estimates a model with a constant and $J - 1$ cutpoints.

As $J$ gets large,

$$\hat{\beta}_{\text{orderedprobit}} \to \hat{\beta}_{\text{OLS}}$$

in finite samples.

Substitute $\Lambda$ for $\Phi$ and you have an ordered logit model.
3 Censored Data

Suppose our latent model is \( y^*_i = x'_i \beta + \varepsilon_i \). However, the data are censored at zero:

\[
\begin{align*}
  y_i &= 0 & \text{if } y^*_i &\leq 0 \\
  y_i &= y^*_i & \text{if } y^*_i &> 0
\end{align*}
\]

(Examples)

Note that for \( \varepsilon \sim N(0, \sigma^2) \), this latent variable model is sort of “half probit, half OLS.” Further, because we partially observe the magnitude of \( y^*_i \), we can recover estimates of \( \sigma^2 \).

If \( y_i = 0 \), \( y^*_i \leq 0 \), which implies

\[
\begin{align*}
  x'_i \beta + \varepsilon_i &\leq 0 \\
  \varepsilon_i &\leq -x'_i \beta
\end{align*}
\]

which occurs with probability \( \Phi(-x'_i \beta / \sigma) \).

If \( y_i > 0 \), we’re in normal regression land, with

\[
f(y_i | x_i, \beta, \sigma^2) = \phi \left( \frac{y_i - x'_i \beta}{\sigma} \right).
\]

Let \( \delta_i = 1_{y_i > 0} \). The likelihood for the full sample is

\[
\prod_{i=1}^N \left[ \phi \left( \frac{y_i - x'_i \beta}{\sigma} \right) \right]^{\delta_i} \left[ \Phi \left( \frac{-x'_i \beta}{\sigma} \right) \right]^{1-\delta_i},
\]

while the log-likelihood is

\[
\sum_{i=1}^N \left\{ \delta_i \ln \left[ \phi \left( \frac{y_i - x'_i \beta}{\sigma} \right) \right] + (1 - \delta_i) \ln \left[ \Phi \left( \frac{-x'_i \beta}{\sigma} \right) \right] \right\}.
\]

This model is called “Tobin’s probit,” or Tobit, for James Tobin (1958).

Note two things: First, there is no reason the censoring point must be zero. Second, there is no reason the data can’t be right-censored instead of left-censored. Duration data are typically right censored. Third, some data have two-limits instead of one. Fourth, the limits might vary from observation to observation.

4 Count Models: The Poisson

Let’s look closely at an important statistical distribution, how it derives from a theory of a data generating process, and how to derive a regression model from it.
Recall the binomial distribution, which represents the probability of \( k \) successes in \( n \) independent trials, where the probability of success in each trial is \( p \). If \( y \sim \text{Binomial}(n, p) \), its mass function is
\[
P(y = k) = \binom{n}{k} p^k (1 - p)^{n-k}.
\]

Think of the binomial in terms of time: Divide a span of time of length one into \( n \) “sub-intervals” of length \( 1/n \). In each, we’ll conduct an experiment with success probability \( p \). (Example, \( n = 5, p = 0.8 \)).

Now, imagine we divide the span into more intervals and make the success probability proportionally smaller:

- \( n = 10 \) and \( p = 0.4 \).
- \( n = 20 \) and \( p = 0.2 \).
- \( n = 40 \) and \( p = 0.1 \).
- \( n = 80 \) and \( p = 0.05 \).

Notice in the example that \( \bar{y} = np = 4 \). Of course, we could use any value \( \lambda = np > 0 \), so long as when we increase the number of trials, \( p \) gets proportionally smaller.

Question: What happens when \( n \to \infty \): an infinite number of trials with miniscule success probability?

At that point, the time span becomes a continuum. Formally, we’re interested in
\[
\lim_{n \to \infty} P(y = k \mid \frac{\lambda}{n}, n).
\]
This equals
\[
\lim_{n \to \infty} \binom{n}{k} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{n-k} = \frac{e^{-\lambda} \lambda^k}{k!}
\]
So
\[
P_{\text{Poisson}}(y = k \mid \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}.
\]
Note that \( E[y] = \lambda \), which is reassuring, because \( E[y] \) given \( y \sim \text{Binomial}(n, p) \) is \( np \), and \( \lambda = np \). \( \lambda \) is often called the rate parameter, recalling the time-based intuition. With the Poisson, we have discrete events occurring in continuous time. Further, since there are an infinite number of “trials,” the support of the Poisson distribution is the non-negative integers.

Note also that
\[
\text{Var}[y] = E[y] = \lambda.
\]
(What does this imply?)
5 The Poisson likelihood

Suppose \( y \) is a vector of draws from a Poisson distribution with mean parameter \( \lambda \). Then

\[
L(y|\lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!},
\]

and the log-likelihood is

\[
\ell(y|\lambda) = \sum_{i=1}^{N} \left[ -\lambda + y_i \ln \lambda - \ln(y_i!) \right]
\]

Now suppose we are interested in a more complicated model of the form (using King’s notation):

\[
Y_i \sim \text{Poisson}(\lambda_i)
\]

\[
\lambda_i = g(x_i, \beta)
\]

We would like \( g(\cdot) \) to be increasing in \( \beta \) and always positive. Typically, we use \( \lambda_i = g(x_i, \beta) = \exp(x_i'\beta) \). Then

\[
\ell(y|x_i, \beta) = \sum_{i=1}^{N} \left[ -\exp(x_i'\beta) + y_i x_i'\beta - \ln(y_i!) \right]
\]

This is called the Poisson regression model. (Advantages and disadvantages)

6 The Negative Binomial Model (Hard)

The foregoing approach suggests \( \lambda_i \) is fully determined by a linear combination of \( x \)s. Suppose, however, it is determined by the \( x \)s and some unobservable, individual-specific random effect, \( \varepsilon_i \), which we assume is uncorrelated with the \( x \)s. Now we have

\[
y_i \sim \text{Poisson}(\lambda_i)
\]

\[
\lambda_i = \exp(x_i'\beta + \varepsilon_i)
\]

If we think of \( \varepsilon \) as something like an OLS error term, then our model has two stochastic components: the process that generates \( \varepsilon \)s and the process that generates \( y \)s conditional on \( \varepsilon \) and our (fixed) \( x_i'\beta \). This is a simple, “hierarchical” model.

Note that we can express

\[
\lambda_i = \exp(x_i'\beta + \varepsilon_i)
\]

\[
= \exp(x_i'\beta) \exp(\varepsilon_i)
\]

\[
= \nu_i \exp(x_i'\beta),
\]
where \( \nu_i > 0 \) by construction. The pdf for \( y_i \) is

\[
f(y_i|x_i, \beta, \nu_i) = \frac{e^{-\nu_i \exp(x'_i \beta)} \nu_i^{y_i} \exp(x'_i \beta)^{y_i}}{y_i!}.
\]

But \( f \) is now a function of this unobservable random variable.

Suppose we wanted to calculate the expected value of \( f \) by averaging over possible values of \( \nu_i \). If the pdf of \( \nu_i \) is \( h(\nu_i) \), then using our formula for the expected value of a function of a random variable, we have:

\[
E_{\nu_i}[f(y_i|x_i, \beta, \nu_i)] = \bar{f}(y_i|x_i, \beta) = \int_{\mathbb{R}^+} f(y_i|x_i, \beta, \nu_i) h(\nu_i) d\nu_i.
\]

7 Time Out: The Gamma distribution

The gamma distribution is a continuous probability distribution with positive support. Its pdf is:

\[
h(t|\theta, \alpha) = \frac{\theta^\alpha t^{\alpha-1} e^{-\theta t}}{\Gamma(\alpha)},
\]

where \( E(t) = \alpha/\theta \), \( Var(t) = \alpha/\theta^2 \).

\( \Gamma(\alpha) \) is called the “Gamma integral,” sort of a continuous version of the factorial function. Note that \( k! = \Gamma(k+1) \).

If we normalize \( \alpha = \theta \), then our gamma r.v. has a mean of one.

8 Time In

Suppose our multiplicative random effect, \( \nu_i \), is distributed unit-mean gamma. Integrating over this distribution to arrive at our expected pdf \( \bar{f} \), we get \( \bar{f}(y_i|x_i, \beta, \alpha) = \)

\[
\frac{\Gamma(y + \alpha)}{\Gamma(\alpha) y!} \left( \frac{\exp(x'_i \beta)}{\exp(x'_i \beta) + \alpha} \right)^y \left( \frac{\alpha}{\exp(x'_i \beta) + \alpha} \right)^\alpha
\]

This horrendous beast is called the Negative binomial density, and can easily be incorporated into a likelihood function. It has

\[
E[y_i] = \exp(x'_i \beta)
\]

\[
Var(y_i) = \exp(x'_i \beta) + \exp(2x'_i \beta)/\alpha
\]

Why in the world would we want to use this as the basis of a statistical model?