Lecture: Limited Dependent Variable Model (Wooldridge’s book chapter 17)
Big Picture

Limited Dependent Variable (LDV) Model can be used when the dependent variable is special. Examples are

1. $y$ is binary (eg: voting for Trump or Biden)—probit or logistic (logit) regression
2. $y$ is categorical with more than two unordered outcomes (eg: drinking Coke, Pepsi, or water)—multinomial logistic regression
3. $y$ is categorical with ordered outcomes (eg: evaluation is above, equal or below average)—ordered logistic regression
4. $y$ represents corner solution (eg: consumption of cigarettes)—Tobit model
5. $y$ represents counts (eg: the number of car accidents)—Poisson model
6. $y$ represents duration (eg: survival time of a patient)—Cox model

In general, LDV model is estimated by maximum likelihood (ML) method.
Binomial Distribution

1. Suppose we flip a coin once. The probability of seeing head is $p$; and the probability of seeing tail is $1 - p$

2. Suppose we flip the same coin twice. The probability of seeing one head and one tail is $(p)(1 - p) + (1 - p)(p) = C_2^1(p)(1 - p)$; the probability of seeing two heads is $C_2^2p^2(1 - p)^0$; the probability of seeing two tails is $C_2^0p^0(1 - p)^2$

3. In general, if we flip the coin $n$ times, the probability of seeing $r$ heads is

$$C_n^r p^r (1 - p)^{n-r} \quad (1)$$

where

$$C_n^r \equiv \binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (2)$$

denotes the binomial coefficient, or combinations of $r$ successes out of $n$ trails.

4. Notice that $C_n^r$ is free of $p$
Joint Probability of IID Sample of Bernoulli Variables

1. When \( n = 1 \), the Binomial distribution reduces to Bernoulli distribution

2. Bernoulli random variable is binary (dichotomous), and its distribution is

\[
y = \begin{cases} 
1 \text{ (success), with probability } p \\
0 \text{ (failure), with probability } 1 - p 
\end{cases}
\]  

(3)

3. Equivalently, we can write the distribution function as

\[
f = p^y (1 - p)^{1-y}, \quad (y = 1, 0)
\]  

(4)

4. For a given iid sample of \((y_1, y_2, \ldots, y_n)\), the joint probability of observing this particular sequence is

\[
\prod_{i=1}^{n} f_i \equiv f_1 f_2 \ldots f_n = (p^{y_1} (1 - p)^{1-y_1}) \ldots (p^{y_n} (1 - p)^{1-y_n}) = p^{\sum y_i} (1 - p)^{n-\sum y_i}
\]

(5)

5. Compared to (1), equation (5) drops \( C_n^r \) since the sample is given (only one combination). Note that \( \sum y_i = r \) if there are \( r \) successes.
Maximum Likelihood (ML) Method

1. The joint probability (5) is a function of \( p \), called **likelihood** function

\[
L \equiv p^{\sum y_i} (1 - p)^{n - \sum y_i}
\]  

(6)

2. When \( p \) is unknown (eg, we are not sure whether the coin is fair) the ML method estimates \( p \) by maximizing the likelihood function (i.e., maximizing the probability of observing the given sample)

\[
\hat{p} = \arg\max L = \arg\max p^{\sum y_i} (1 - p)^{n - \sum y_i}
\]

(7)

3. In practice, ML maximizes the log likelihood since the log transformation is monotonic

\[
\hat{p} = \arg\max \log(L) = \arg\max (\sum y_i) \log(p) + (n - \sum y_i) \log(1 - p)
\]

(8)

4. Taking derivative with respect to \( p \) and setting to zero, we have

\[
\hat{p} = \frac{\sum y_i}{n} = \frac{r}{n} = \text{sample proportion}
\]

(9)

Thus sample proportion is the ML estimate for population proportion
Linear Probability Model (LPM)

1. So far we assume $p$ is constant since we flip the same coin again and again. In reality, we ask different persons whether voting for Trump or Biden, so $p$ is no longer constant.

2. We want to use covariates $x$ such as gender, age, education to explain $p$. LPM is based on a linear function

$$ p_i \equiv P(y_i = 1|x_i) = E(y_i|x_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} $$ (10)

where we use the fact that for a Bernoulli random variable, $P(y = 1) = E(y)$

3. (Critical thinking) can $x$ include choice-specific characteristics such as age of the presidential candidate, or price of drinks?

4. Basically equation (10) is the population regression function (conditional mean). Thus LPM amounts to regressing $y$ onto $x$ via OLS

5. Heteroskedasticity-robust standard error should be used for LPM since

$$ \text{var}(y_i|x_i) = p_i(1-p_i) \neq \text{constant} $$ (11)
Probit Model

1. LPM has low computational cost and constant marginal effect, but it fails to account for two facts that (i) $y$ is special (binary), and (ii) probability should be bounded $0 \leq p_i \leq 1$.

2. Those drawbacks of LPM motivate nonlinear models such as Probit and Logit models.

3. Probit model uses the cumulative distribution function (cdf) of normal distribution $\Phi$ to specify the probability, and $0 \leq \Phi \leq 1$ by construction

$$p_i = \Phi(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}) = \int_{-\infty}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$ (12)

4. Now the unknown parameters are $\beta$s, and ML estimates them by maximizing log($L$)

$$\hat{\beta} = \arg\max \log(\Pi_{i=1}^n f_i) = \sum_{i=1}^n \log(f_i) = \sum_{i=1}^n \log(p_i^{y_i}(1-p_i)^{1-y_i}) = \sum_{i=1}^n [y_i \log(p_i) + (1-y_i) \log(1-p_i)]$$ (13)

Numerical methods are used to solve this optimization problem since there is no closed-form or analytical solution.
Derivation of Probit Model using Latent Variable

1. It seems that the $\Phi$ thing comes from nowhere. Well, we can use a latent variable to justify it.

2. Suppose a voter gives each presidential candidate a score, and the score depends on the voter’s gender, age, educ, etc. The score is unobserved, so is called latent variable. A person votes for Trump, $y = 1$, only if his score is above a threshold, say, 0

\[
P(y = 1) = P(score > 0) \tag{14}
\]
\[
= P(x\beta + u > 0) \tag{15}
\]
\[
= P(u > -x\beta) \tag{16}
\]
\[
= P(u < x\beta) \tag{17}
\]
\[
\equiv \Phi(x\beta) \tag{18}
\]

where we assume the unobserved factors $u$ follow standard normal distribution, and we use the fact that normal distribution is symmetric. The last equality is the definition for cdf. No wonder $0 \leq \Phi \leq 1$

3. (Critical thinking) what if the threshold is not 0?
Marginal Effect

1. Probit model has the benefit of imposing the 0-1 boundary for probability. The cost is that $\beta$ alone does not measure the (magnitude of) marginal effect.

2. We can apply chain rule to show the marginal effect of the $j$-th covariate on probability is given by

$$
\frac{\partial p_i}{\partial x_{ji}} = \phi(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}) \beta_j
$$

where $\phi$ is the derivative of $\Phi$, called probability density function (pdf). In short, the marginal effect of probit model is non-constant.

3. $\phi$ is non-negative. So the sign of marginal effect is determined by $\beta$

4. Because $x$ varies, in practice, we can compute the average marginal effect (AME):

$$
AME = \beta_j \frac{\sum_{i=1}^{n} \phi(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki})}{n}
$$

or the marginal effect at average (MEAA):

$$
MEAA = \beta_j (\beta_0 + \beta_1 \bar{x}_1 + \ldots + \beta_k \bar{x}_k)
$$
Logit Model (Logistic Regression)

1. Probit model has the drawback that $\Phi$ and $\phi$ are hard to compute

2. Alternatively, one may use a Logit model that specifies the success probability as

$$p_i = \Lambda(x_i\beta) \equiv \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}$$

where $\Lambda$ denotes the cdf of a logistic distribution. We can directly verify that $0 \leq \Lambda \leq 1$

3. Like the probit model, the marginal effect is product of $\beta$ and a factor

$$\frac{\partial p_i}{\partial x_i} = \Lambda'(x_i\beta)\beta$$

where $\Lambda'$ denotes the derivative of $\Lambda$. The sign of marginal effect only depends on the sign of $\beta$.

4. Logistic distribution looks similar to normal distribution. So in general, probit and logit models produce similar marginal effects
Odds, Log Odds

1. In industry logit model is more popular than probit model because of a simple formula for odds, and interpretation of coefficient in terms of log odds

2. By definition

\[
\text{odds} \equiv \frac{p_i}{1 - p_i} = e^{x_i \beta} \tag{24}
\]

\[
\log \text{ odds} = x_i \beta \tag{25}
\]

So people interpret $\beta$ as the effect on log odds when $x$ changes by one unit.

3. Note that even though $p_i$ is bounded between 0 and 1, the log odds is unbounded

4. Equation (25) is an example of generalized linear model (GLM) in which the right hand side is an unrestricted linear function, but the left hand side is a transformation of original data
Odds Ratio

1. We are interested in a special case where $x$ is a dummy variable. It follows that

\[
\text{odds ratio} \equiv \frac{\text{odds when } x = 1}{\text{odds when } x = 0} = e^{\beta}
\]  

(26)

So the exponential of coefficient of a dummy independent variable in a logit model gives odds ratio

2. For instance, if odds ratio $= 2$, that means the odds when $x = 1$ is twice of odds when $x = 0$

3. We can test the null hypothesis that $x$ does not matter—$H_0 : \text{odds ratio} = 1$, which is the same as $H_0 : \beta = 0$,  