

Notes on Discrete Choice/Limited and Qualitative Dependent Variables

Ronald L. Oaxaca

University of Arizona

August 2011

I. BACKGROUND

The models to be presented here are ones in which the range of the dependent variable is limited in some way. The variable's range may be confined to discrete values or to a mixture of discrete and continuous values. These models are very common in applied work and often involve special estimation methods. The training course will cover the essentials of estimating empirical models in which the dependent variable is either discrete or bounded.

We will begin with the simplest models in which the dependent variable is binary, e.g. 0 or 1. These models include the Linear Probability Model, the Probit model, and the Logit model. Such models will then be extended to include more than two outcomes or choices. The multinomial logit model is commonly used to estimate the effects of variables on the probabilities of each of the mutually exclusive outcomes. If the outcomes are ordered in some way, then ordered probit/logit is more appropriate for estimating the probabilities.

Another class of models is appropriate when the range of the dependent variable is limited in some fashion. This class of limited dependent variable models includes truncated regressions, censored regressions (Tobit), and selectivity correction (Heckit) models. These models often assume that the underlying error distribution is a normal.

All standard statistical/econometrics programs have commands to estimate the models covered here and include estimation of the marginal effects, standard errors, t values, and other summary statistics.

II. DISCRETE CHOICE/QUALITATIVE LIMITED DEPENDENT VARIABLES

Background Reading:

Greene W.H., *Econometric Analysis*, 6th ed., Chapter 23 (pp. 770-793, 813-817, 831-835, 841-847).

Wooldridge, J.M., *Introductory Econometrics: A Modern Approach*, 4th ed., Chapter 7 (pp. 246-251), Chapter 17 (pp. 574-587)

A. Linear Probability Model

Consider a binary choice or outcome that is measured by a variable Y_i from a random sample where

$$\begin{aligned} Y_i &= 1 \text{ if some event occurs or choice is made} \\ &= 0 \text{ otherwise.} \end{aligned}$$

The probability that the choice or event will occur for any observation from a random sample can in general be expressed as $P_i = \text{prob}(Y_i = 1|I_i)$, where

$$\text{prob}(Y_i = 1|I_i) = F(I_i), \quad i = 1, \dots, n.$$

The function $F(I_i)$ is a cumulative distribution function (cdf), I_i is an index function such that

$$\begin{aligned} I_i &= X_i' \beta \\ &= \beta_0 + \sum_{k=1}^K \beta_k X_{ik} \end{aligned}$$

where X'_i is a $1 \times k$ observation vector, β is a $k \times 1$ conforming parameter vector, and X_{ik} is the k th explanatory variable for observation i . An example might be $Y_i = 1$ if individual i is in the labor force and $= 0$ otherwise (not in the labor force). In this example the X_{ik} variables might include marital status, age, education, etc. Note that the probability that the individual is not in the labor force is

$$\begin{aligned} \text{prob}(Y_i = 0|I_i) &= 1 - P_i \\ &= 1 - \text{prob}(Y_i = 1|I_i) \\ &= 1 - F(I_i). \end{aligned}$$

The simplest discrete choice or binary outcome model is the Linear Probability Model (LPM). The LPM is specified by

$$\begin{aligned} Y_i &= I_i + \varepsilon_i \\ &= X'_i \beta + \varepsilon_i \\ &= \beta_0 + \sum_{k=1}^K \beta_k X_{ik} + \varepsilon_i \end{aligned}$$

where ε_i is a random error term. Note

$$\begin{aligned} \varepsilon_i &= 1 - I_i \text{ if } Y_i = 1 \\ &= -I_i \text{ if } Y_i = 0. \end{aligned}$$

The distribution of ε_i is discrete and may be expressed as

ε_i	$f(\varepsilon_i)$
$1 - I_i$	P_i
$-I_i$	$1 - P_i$

A standard assumption is that the population mean of the error term is 0, i.e.

$E(\varepsilon_i) = 0 \forall i$. We can use this assumption to identify P_i :

$$E(\varepsilon_i) = (1 - I_i)(P_i) + (-I_i)(1 - P_i) = 0 \Rightarrow P_i = I_i = X'_i \beta = \beta_0 + \sum_{k=1}^K \beta_k X_{ik}.$$

The marginal effects of the X 's are given by β_k where $\beta_k = \frac{\partial Y}{\partial X_k}$ if X_k is a continuous variable and $\beta_k = \Delta Y = Y_{(X_k=1)} - Y_{(X_k=0)}$ if X_k is a binary (1,0) variable.

There are some well known problems with estimation of the LPM by ordinary least squares (*OLS*). First of all consider the variance of the error term ε_i . We can determine the variance of ε_i from the discrete distribution table given above.

$$\begin{aligned} \text{Var}(\varepsilon_i) &= (1 - I_i)^2 (P_i) + (-I_i)^2 (1 - P_i) \\ &= (1 - I_i)^2 (I_i) + (-I_i)^2 (1 - I_i) \\ &= (1 - I_i) (I_i) \\ &= (1 - P_i) (P_i). \end{aligned}$$

The problem here is one of heteroscedasticity in that the variance of the error term is not constant as it depends on I_i which in turn depends on the X'_{ik} s. Since the variance is different for each observation, the observations are not equally reliable. Strictly speaking, the ε'_i s do not constitute a random sample because each error term comes from a different distribution (same mean but different variances). The estimated standard errors and t values on the β' s are not valid for statistical inference. One could use techniques to try to correct for the heteroscedasticity, e.g. the White correction procedure.

A second problem with *OLS* estimation of the LPM has to do with the fact that although $0 \leq P_i \leq 1$, there is no guarantee that $0 \leq \hat{P}_i \leq 1$ where $\hat{P}_i = \hat{I}_i = X'_i \beta$. One might seek to use advanced estimation methods to overcome both the heteroscedasticity problem and the problem with bounding the probability estimate \hat{P}_i between 0 and 1.

Even if *OLS* is not an appropriate estimator of the LPM, the LPM itself can be appropriate in cases in which the probability of an outcome or choice can literally be

0 or 1. So the model might be specified as

$$\begin{aligned} Y_i &= 1 (\varepsilon_i = 0) \text{ if } I_i \geq 1 \\ &= 0 (\varepsilon_i = 0) \text{ if } I_i \leq 0 \\ &= I_i + \varepsilon_i \text{ if } 0 < I_i < 1. \end{aligned}$$

For example, imagine a job promotion process in which $I_i = \beta_0 + \sum_{k=1}^K \beta_k X_{ik}$ is an index of job performance. The X variables are various job performance measures, e.g. absentee rate, seniority, education, and the β 's are the weights placed on the performance measures. Suppose a person will definitely be promoted if $I_i \geq 1$ and will definitely not be promoted if $I_i \leq 0$. For all others, the probability of promotion is $P_i = I_i = \beta_0 + \sum_{k=1}^K \beta_j X_{ik}$, for $0 < I_i < 1$.

B. Probit Model

The probit model might be used in a situation in which the probability of a choice or outcome can never literally be 0 or 1. Think of the following linear, latent variable model:

$$Y_i^* = I_i + \varepsilon_i$$

where ε_i is independently and identically distributed (IID) $N(0, \sigma_\varepsilon^2)$. However, Y_i^* is not observed but its relationship to the X 's in the model can be inferred from the following information on a binary variable Y_i which is observed:

$$\begin{aligned} Y_i &= 1 \text{ if } Y_i^* > 0 \\ &= 0 \text{ if } Y_i^* \leq 0. \end{aligned}$$

Variables such as Y_i^* are known as latent variables. An example might be the utility of a choice. So Y_i^* might be the utility of choosing to participate in the labor force. Let $Y_i = 1$ if the i th individual is in the labor force, and = 0 otherwise. For

people who are observed to be in the labor force, $Y_i = 1$, and the presumption is that $Y_i^* > 0$. Likewise, for someone observed to be out of the labor force, $Y_i = 0$, and the presumption is that $Y_i^* \leq 0$.

We can set up the probit model in the following way:

$$\begin{aligned} Y_i &= 1 \text{ if } I_i + \varepsilon_i > 0 \text{ (or } \varepsilon_i > -I_i) \\ &= 0 \text{ if } I_i + \varepsilon_i \leq 0 \text{ (or } \varepsilon_i \leq -I_i). \end{aligned}$$

Because σ_ε^2 usually cannot be identified in the probit model, it is normalized to 1 so that one assumes $\varepsilon_i \sim N(0, 1)$. In effect the original β' s in the model are being divided by the unobserved standard deviation σ_ε . The choice or outcome probabilities are defined according to

$$\begin{aligned} \text{prob}(Y_i = 1|I_i) &= \text{prob}(\varepsilon_i > -I_i) \\ &= \text{prob}(\varepsilon_i < I_i) \text{ (symmetry)} \\ &= \Phi(I_i) \end{aligned}$$

where $\Phi(I_i) = F(I_i)$ is the cdf for the standard normal distribution. It follows that $\text{prob}(Y_i = 0|I_i) = 1 - \Phi(I_i)$.

Estimation of the β' s are obtained by maximizing the likelihood function for the sample:

$$L = \prod_{i=1}^n [\Phi(I_i)]^{y_i} [(1 - \Phi(I_i))]^{1-y_i}$$

where $y_i = 0, 1$ is the actual binary outcome for the random variable Y_i . It is generally easier to work in terms of the log likelihood function:

$$\ln(L) = \sum_{i=1}^n \{y_i \ln [\Phi(I_i)] + (1 - y_i) \ln [(1 - \Phi(I_i))]\}.$$

Note that in the probit model, the β' s are not the marginal effects of the X variables on the probability $P_i = \text{prob}(Y_i = 1|I_i)$. The reason is that when there is a change in

one of the X variables, this affects the index function I_i , which affects the cdf $\Phi(I_i)$, which in turn affects $\text{prob}(Y_i = 1|I_i)$. In the case of the probit model, the marginal effects are given by

$$\frac{\partial P_i}{\partial X_{ik}} = \beta_k \phi(I_i)$$

for a continuous variable X_j , where $\phi(I_i) = \frac{\partial \Phi(I_i)}{\partial I_i}$ is the standard normal density function. The marginal effect of a binary variable X_k on the probability P_i is simply $\Delta P_{ik} = \Phi(I_{i(X_{ik}=1)}) - \Phi(I_{i(X_{ik}=0)})$.

How does one estimate the sample-wide marginal effects? One possibility is to obtain the estimate

$$\widehat{\frac{\partial P}{\partial X_k}} = \hat{\beta}_k \phi(\bar{I})$$

for X_k a continuous variable, where $\bar{I} = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k \bar{X}_k$ and the \bar{X}_k variables are the sample means. In the case of a binary X_k variable, we have

$$\widehat{\Delta P_k} = \Phi(\bar{I}_{(X_k=1)}) - \Phi(\bar{I}_{(X_k=0)}).$$

This method evaluates everything at the sample mean. An equally valid alternative is to calculate the sample average of all individual's marginal effects:

$$\overline{\frac{\partial P}{\partial X_k}} = \frac{\sum_{i=1}^n \hat{\beta}_k \phi(\hat{I}_i)}{n}$$

for X_k a continuous variable and $\hat{I}_i = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k X_{ik}$. In the case of a binary X_k variable, we have

$$\overline{\Delta P_k} = \frac{\sum_{i=1}^n \widehat{\Delta P_{ik}}}{n},$$

where $\widehat{\Delta P_{ik}} = \Phi(\hat{I}_{i(X_k=1)}) - \Phi(\hat{I}_{i(X_k=0)})$. Typically, any software program that estimates probit models will also report the estimated marginal effects according to either method.

It is perhaps interesting and useful to know that in the probit model, the average of the predicted probabilities for the sample will not generally equal the sample mean proportion of 1's, i.e. $\bar{P} = \frac{\sum_{i=1}^n \hat{P}_i}{n} \neq \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$, where $\hat{P}_i = \Phi(\hat{I}_i)$.

Standard econometrics software report estimated standard errors and t values for both the $\hat{\beta}'$ s and the marginal effects. In terms of deciding whether or not a particular X_k variable belongs in the model, it is the estimated standard error or t value associated with $\hat{\beta}_k$ that matters, not the estimated standard error or t value associated with the marginal effect of X_k .

Goodness of fit measures for the probit model are not quite as straightforward as the R^2 for the linear regression model. One common measure for the probit model is the McFadden's pseudo $R^2 = 1 - \frac{\ln(L_{MLE})}{\ln(L_0)}$, where $\ln(L_{MLE})$ is the value of the log of the likelihood function evaluated at the maximum likelihood estimates of the β' s, and $\ln(L_0)$ is the value of the log of the likelihood function when the slope coefficients are set equal to 0, i.e. $\hat{\beta}_k = 0$ for $k = 1, \dots, K$. This statistic is sometimes referred to as the Likelihood Ratio Index (LRI). It turns out that $0 \leq LRI < 1$. A value of 0 for the pseudo R^2 indicates that none of the X_k variables have any explanatory value. In some sense the higher the value of the pseudo R^2 , the better the fit. Unfortunately, there is no convenient interpretation of the pseudo R^2 as there is with the conventional R^2 from a standard regression model.

Another measure of fit is the predictive accuracy rate. The model can be used to predict the choice or outcome according to

$$\begin{aligned} \hat{Y}_i &= 1 \text{ if } \Phi(\hat{I}_i) > 0.5 \\ &= 0 \text{ if } \Phi(\hat{I}_i) \leq 0.5. \end{aligned}$$

Consider the following definitions:

$$\begin{aligned}
 N_{1.} &= \text{the actual number of 1's} \\
 N_{0.} &= \text{the actual number of 0's} \\
 N_{.1} &= \text{the predicted number of 1's} \\
 N_{.0} &= \text{the predicted number of 0's.}
 \end{aligned}$$

We construct a prediction table according to

		Predicted	
		Y = 0	Y = 1
Actual	Y = 0	N_{00}	N_{01}
	Y = 1	N_{10}	N_{11}

where N_{00} = the number of 0's that were correctly predicted to be 0, N_{01} = the number of 0's that were incorrectly predicted to be 1, N_{10} = the number of 1's that were incorrectly predicted to be 0, and N_{11} = the number of 1's that were correctly predicted to be 1. It should be clear that $N_{00} + N_{01} = N_{0.}$ (the actual number of 0's), $N_{10} + N_{11} = N_{1.}$ (the actual number of 1's), $N_{00} + N_{10} = N_{.0}$ (the predicted number of 0's), and $N_{01} + N_{11} = N_{.1}$ (the predicted number of 1's). A common measure of goodness of fit is the success rate $\frac{N_{00} + N_{11}}{n}$, where $n = N_{00} + N_{01} + N_{10} + N_{11}$ (total sample size). One problem with the success rate measure is that if the sample is lopsided, e.g., $\frac{N_{1.}}{n} = 0.9$ (the event or choice occurred for 90% of the sample), one could achieve an accuracy rate of 90% by simply predicting the event or choice will occur for every observation without ever estimating the model! In fact if $\Phi(\hat{I}_i) > 0.5$ for every observation, $\hat{Y}_i = 1$ for every observation. In this case the accuracy rate $= \frac{N_{1.}}{n} > 0.5$.

C. Logit Model

The logit model is virtually identical to the probit model except for the assumed distribution of the error term ε_i . In the case of the logit model, the error term follows a logistic distribution with mean 0 and variance $\frac{\pi^2 s^2}{3}$, where s is a scale parameter. Similar to the probit model, we have

$$\begin{aligned} Y_i &= 1 \text{ if } I_i + \varepsilon_i > 0 \text{ (or } \varepsilon_i > -I_i) \\ &= 0 \text{ if } I_i + \varepsilon_i \leq 0 \text{ (or } \varepsilon_i \leq -I_i). \end{aligned}$$

Because the scale parameter s is generally not identified in the logit model, it is normalized to 1. Consequently σ_ε^2 is normalized to $\frac{\pi^2}{3}$ so that one assumes $\varepsilon_i \sim$ logistic with mean 0 and variance $\frac{\pi^2}{3}$. The choice or outcome probabilities are defined according to

$$\begin{aligned} \text{prob}(Y_i = 1|I_i) &= \text{prob}(\varepsilon_i > -I_i) \\ &= \text{prob}(\varepsilon_i < I_i) \text{ (symmetry)} \\ &= \Lambda(I_i) \end{aligned}$$

where $\Lambda(I_i)$ is the cdf for the logistic distribution. The logistic cdf is defined by

$$\begin{aligned} \Lambda(I_i) &= \frac{e^{I_i}}{1 + e^{I_i}} \\ &= \frac{1}{1 + e^{-I_i}}. \end{aligned}$$

It follows that

$$\begin{aligned} \text{prob}(Y_i = 0|I_i) &= 1 - \Lambda(I_i) \\ &= \frac{e^{-I_i}}{1 + e^{-I_i}} \\ &= \frac{1}{1 + e^{I_i}}. \end{aligned}$$

Other useful relationships for the logistic distribution include the density function $\lambda(I_i) = \frac{\partial \Lambda(I_i)}{\partial I_i}$, where

$$\begin{aligned}\lambda(I_i) &= \frac{e^{-I_i}}{(1 + e^{-I_i})^2} \\ &= [1 - \Lambda(I_i)] [\Lambda(I_i)].\end{aligned}$$

Later on with Tobit and Heckit models, the Inverse Mills Ratio (IMR) and hazard rate will be of particular interest. The IMR is the ratio of the density function to the cdf. In the case of the logit model

$$\begin{aligned}\text{IMR} &= \frac{\lambda(I_i)}{\Lambda(I_i)} \\ &= 1 - \Lambda(I_i).\end{aligned}$$

The hazard rate is the ratio of the density to 1 - cdf. In the logit case we have

$$\begin{aligned}\text{hazard rate} &= \frac{\lambda(I_i)}{1 - \Lambda(I_i)} \\ &= \Lambda(I_i).\end{aligned}$$

Estimates of the β' s in $I_i = \beta_0 + \sum_{k=1}^K \beta_j X_{ik}$ are obtained by maximizing the likelihood function or the log likelihood function for the sample:

$$\begin{aligned}L &= \prod_{i=1}^n [(\Lambda(I_i))^{y_i} [(1 - \Lambda(I_i))]^{1-y_i}] \\ \ln(L) &= \sum_{i=1}^n \{y_i \ln [\Lambda(I_i)] + (1 - y_i) \ln [(1 - \Lambda(I_i))]\},\end{aligned}$$

where $y_i = 0, 1$ is the actual binary outcome for the random variable Y_i .

Note that as in the probit model, the β' s are not the marginal effects of the X variables on the probability $P_i = \text{prob}(Y_i = 1|I_i)$. In the case of the logit model, the marginal effects are given by

$$\frac{\partial P_i}{\partial X_{ik}} = \beta_k \lambda(I_i)$$

for a continuous variable X_k , where $\lambda(I_i) = \frac{\partial \Lambda(I_i)}{\partial X_{ik}}$ is the logistic density function. The marginal effect of a binary variable X_k on the probability P_i is simply $\Delta P_{ik} = \Lambda(I_{i(X_{ik}=1)}) - \Lambda(I_{i(X_{ik}=0)})$.

How does one estimate the sample-wide marginal effects? One possibility is to obtain the estimate

$$\frac{\widehat{\partial P}}{\partial X_k} = \hat{\beta}_k \lambda(\bar{I})$$

for X_j a continuous variable, where $\bar{I} = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k \bar{X}_k$ and the \bar{X}_k variables are the sample means. In the case of a binary X_k variable, we have

$$\widehat{\Delta P}_j = \Lambda(\bar{I}_{(X_k=1)}) - \Lambda(\bar{I}_{(X_k=0)}).$$

This method evaluates everything at the sample mean. An equally valid alternative is to calculate the sample average of all individuals' marginal effects:

$$\frac{\overline{\partial P}}{\partial X_k} = \frac{\sum_{i=1}^n \hat{\beta}_k \lambda(\hat{I}_i)}{n}$$

for X_k a continuous variable and $\hat{I}_i = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k X_{ik}$. In the case of a binary X_k variable, we have

$$\overline{\Delta P}_k = \frac{\sum_{i=1}^n \widehat{\Delta P}_{ik}}{n},$$

where $\widehat{\Delta P}_{ik} = \Lambda(\hat{I}_i(X_{ik}=1)) - \Lambda(\hat{I}_i(X_{ik}=0))$. Typically, any software program that estimates logit models will also report the estimated marginal effects according to either method.

It is interesting and useful to know that in the logit model (unlike the probit model), the average of the predicted probabilities for the sample will exactly equal the sample mean proportion of 1's as long as there is a constant term $\hat{\beta}_0$, i.e. $\bar{P} = \frac{\sum_{i=1}^n \hat{P}_i}{n} = \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$, where $\hat{P}_i = \Lambda(\hat{I}_i)$.

Standard econometrics software report estimated standard errors and t values for both the $\hat{\beta}'$ s and the marginal effects. Also, the pseudo R^2 and predictive success rates can be used to evaluate the goodness of fit of the logit model. The logit model can be used to predict the choice or outcome according to

$$\begin{aligned}\hat{Y}_i &= 1 \text{ if } \Lambda(\hat{I}_i) > 0.5 \\ &= 0 \text{ if } \Lambda(\hat{I}_i) \leq 0.5.\end{aligned}$$

III. MULTINOMIAL AND CONDITIONAL LOGIT MODELS

Background Reading:

Greene W.H., *Econometric Analysis*, 6th ed., Chapter 23 (pp. 831-835, 843-850).

McFadden, D., "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed. *Frontiers in Econometrics*. New York, Academic Press, 1974.

A. Multinomial Logit

Suppose there are more than 2 possible choices or outcomes that are mutually exclusive, i.e. only one choice or outcome can occur. To make things concrete, assume that there are $J + 1$ possible outcomes or choices. Furthermore, suppose that the probabilities of these choices depend on the characteristics of the agent making the choice. An example might be to choose among 5 mutually exclusive occupations: professional, clerical, sales, skilled trade, unskilled. The outcomes of these choices might depend on an individual's education, age, etc.

- Let $P_{ij} = \text{prob}(Y_{ij} = 1 | I_{ij})$ represent the probability that agent or individual i selects choice j , where $j = 0, 1, \dots, J$, I_{ij} is the index function corresponding to the j th outcome and is defined by

$$\begin{aligned} I_{ij} &= X_i' \beta_j \\ &= \beta_{0j} + \sum_{k=1}^K \beta_{kj} X_{ik}. \end{aligned}$$

All of the probabilities have to add to 1, i.e. $\sum_{j=0}^J P_{ij} = 1$, and each probability has to be bounded between 0 and 1, i.e. $0 < P_{ij} < 1$.

While there is a multinomial version of the LPM, this is rarely used. There is also a multinomial probit model that can deal with choice or outcomes among several

alternatives. However, the most commonly used model is the multinomial logit model:

$$P_{ij} = \frac{e^{I_{ij}}}{\sum_{j=0}^J e^{I_{ij}}}.$$

The ratios of the probabilities or odds are given by

$$\begin{aligned} \frac{P_{ij}}{P_{i0}} &= \frac{e^{I_{ij}}}{e^{I_{i0}}} \\ &= e^{(I_{ij}-I_{i0})}, \quad j = 1, \dots, J \end{aligned}$$

where $I_{ij} - I_{i0} = (\beta_{0j} - \beta_{00}) + \sum_{k=1}^K (\beta_{kj} - \beta_{k0}) X_{ik}$. Note that only the parameter differences can be identified, so the parameters of the selected base choice/outcome are set to 0, i.e. $\beta_{k0} = 0$ for $k = 0, 1, \dots, K$. This means that $I_{i0} = 0$ and that $\frac{P_{ij}}{P_{i0}} = e^{I_{ij}}$. With this normalization, the individual probabilities may be written as

$$\begin{aligned} P_{i0} &= \frac{1}{1 + \sum_{j=1}^J e^{I_{ij}}} \\ P_{ij} &= \frac{e^{I_{ij}}}{1 + \sum_{j=1}^J e^{I_{ij}}}. \end{aligned}$$

In terms of log odds, the relative probabilities may be expressed as

$$\begin{aligned} \ln \left(\frac{P_{ij}}{P_{i0}} \right) &= I_{ij} \\ &= \beta_{0j} + \sum_{k=1}^K \beta_{kj} X_{ik}, \end{aligned}$$

which is a linear function of the X'_{ik} s.

Estimation of the β' s are obtained by maximizing the likelihood function or log likelihood function for the sample:

$$\begin{aligned} L &= \prod_{i=1}^n P_{i0}^{y_{i0}} P_{i1}^{y_{i1}} \dots P_{iJ}^{y_{iJ}} \\ \ln(L) &= \sum_{i=1}^n \sum_{j=0}^J y_{ij} \ln(P_{ij}), \end{aligned}$$

where $y_{ij} = 0, 1$ is the actual binary outcome for the random variable Y_{ij} .

As with the probit and logit models, the β_{kj} parameters are not the marginal effects of X variables on the probabilities. It turns out that the marginal effect of continuous variable X_k on each probability is given by

$$\begin{aligned}\frac{\partial P_{ij}}{\partial X_{ik}} &= P_{ij} \left(\beta_{kj} - \sum_{m=1}^J \beta_{km} P_{im} \right), \text{ for } j = 1, \dots, J \\ \frac{\partial P_{i0}}{\partial X_{ik}} &= -P_{i0} \left(\sum_{m=1}^J \beta_{km} P_{im} \right), \text{ for } j = 0.\end{aligned}$$

It should be clear that $\sum_{j=0}^J \frac{\partial P_{ij}}{\partial X_{ik}} = 0$ since the effect of a variable on any probability must have an exactly offsetting effect on the other probabilities because the probabilities have to add to 1. The marginal effect of a binary variable X_k on the probability P_{ij} is simply $\Delta P_{ijk} = P_{ij}(I_{ij(X_{ik}=1)}) - P_{ij}(I_{ij(X_{ik}=0)})$.

The estimated probabilities for each observation are obtained from the *MLE* estimated parameters:

$$\begin{aligned}\hat{P}_{i0} &= \frac{1}{1 + \sum_{j=1}^J e^{\hat{I}_{ij}}} \\ \hat{P}_{ij} &= \frac{e^{\hat{I}_{ij}}}{1 + \sum_{j=1}^J e^{\hat{I}_{ij}}},\end{aligned}$$

where $\hat{I}_{ij} = \hat{\beta}_{0j} + \sum_{k=1}^K \hat{\beta}_{kj} X_{ik}$. As long as there is a constant term in the multinomial logit model, the sample average of the predicted probabilities will exactly equal the sample proportions for the outcomes:

$$\begin{aligned}\bar{P}_j &= \frac{\sum_{i=1}^n \hat{P}_{ij}}{n} \\ &= \frac{n_j}{n} \\ &= \bar{Y}_j,\end{aligned}$$

where n_j is the number of observations for which outcome j occurred.

How does one estimate the sample-wide marginal effects? One possibility is to obtain the estimated probabilities at the sample mean of the X 's:

$$\begin{aligned}\hat{P}_0 &= \frac{1}{1 + \sum_{j=1}^J e^{\bar{I}_j}} \\ \hat{P}_j &= \frac{e^{\bar{I}_j}}{1 + \sum_{j=1}^J e^{\bar{I}_j}},\end{aligned}$$

where $\bar{I}_j = \hat{\beta}_{0j} + \sum_{k=1}^K \hat{\beta}_{kj} \bar{X}_k$. The marginal effects would be calculated according to

$$\widehat{\frac{\partial P_j}{\partial X_k}} = \hat{P}_j \left(\hat{\beta}_{kj} - \sum_{m=1}^J \hat{\beta}_{km} \hat{P}_m \right)$$

for X_k a continuous variable. In the case of a binary X_k variable, we have

$$\widehat{\Delta P_{jk}} = \hat{P}_{j(X_k=1)} - \hat{P}_{j(X_k=0)}.$$

An equally valid alternative is to calculate the sample average of all individual's marginal effects:

$$\overline{\frac{\partial P_j}{\partial X_k}} = \frac{\sum_{i=1}^n \left(\widehat{\frac{\partial P_{ij}}{\partial X_{ik}}} \right)}{n}$$

for X_k a continuous variable, $\widehat{\frac{\partial P_{ij}}{\partial X_{ik}}} = \hat{P}_{ij} \left(\hat{\beta}_{kj} - \sum_{m=1}^J \hat{\beta}_{km} \hat{P}_{im} \right)$, and

$\widehat{\frac{\partial P_{i0}}{\partial X_{ik}}} = -\hat{P}_{i0} \left(\sum_{m=1}^J \hat{\beta}_{km} \hat{P}_{im} \right)$. In the case of a binary X_k variable, we have

$$\overline{\Delta P_{jk}} = \frac{\sum_{i=1}^n \widehat{\Delta P_{ijk}}}{n},$$

where $\widehat{\Delta P_{ijk}} = \hat{P}_{ij(X_{ik}=1)} - \hat{P}_{ij(X_{ik}=0)}$.

The pseudo R^2 and predictive success rates can be used to evaluate the goodness of fit of the multinomial logit model. The multinomial logit model can be used to

predict the choice or outcome according to

$$\begin{aligned}\hat{Y}_{ij} &= 1 \text{ if } \hat{P}_{ij} = \max \left\{ \hat{P}_{i0}, \hat{P}_{i1}, \dots, \hat{P}_{iJ} \right\} \\ &= 0 \text{ otherwise.}\end{aligned}$$

The success rate = $\frac{\sum_{j=0}^J N_{jj}}{n}$, where N_{jj} = the number of actual j choices that were correctly predicted to be j choices.

B. Conditional Logit Model

The conditional logit model is sometimes referred to as McFadden's conditional logit model. As in the case of the multinomial logit model, there can be more than 2 choices or outcomes. The difference is that it is the attributes of the choice that determine the probabilities of each choice rather than the characteristics of the chooser or individual agent. So for example, consider the decision of whether to travel by air, rail, bus, or car. The attributes of the choice might consist of the travel time and cost for each travel mode for each individual traveler.

McFadden's original formulation of the model (1974) was as a random utility model:

$$U_{ij} = I_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, n \text{ (observations)}, \quad j = 1, \dots, J \text{ (alternatives)},$$

where I_{ij} is the index function defined by

$$\begin{aligned}I_{ij} &= Z'_{ij}\beta \\ &= \sum_{k=1}^K \beta_k Z_{ijk}\end{aligned}$$

where Z_{ijk} is the k th attribute of the j th alternative for individual i , and ε_{ij} is a random error that follows the extreme value distribution. Notice that the β parameters are the same for every choice because they reflect the effects of each choice attribute on the utility of the choice. The idea here is that the utility of an observed choice

for an individual must have been greater than the utility from any other choice the individual could have made. So for example, consider the probability of individual i choosing alternative 1:

$$\begin{aligned} \text{prob}(Y_{i1} = 1) &= \text{prob}(U_{i2} - U_{i1} < 0, U_{i3} - U_{i1} < 0, \dots, U_{iJ} - U_{i1} < 0) \\ &= \text{prob}(\varepsilon_{i2} - \varepsilon_{i1} < I_{i1} - I_{i2}, \varepsilon_{i3} - \varepsilon_{i1} < I_{i1} - I_{i3}, \dots, \varepsilon_{iJ} - \varepsilon_{i1} < I_{i1} - I_{iJ}). \end{aligned}$$

It turns out that $\varepsilon_{ij} - \varepsilon_{im} \sim \text{logistic}$ for $m \neq j$.

The probabilities are given by

$$\begin{aligned} P_{ij} &= \text{prob}(Y_{ij} = 1) \\ &= \frac{1}{\sum_{m=1}^J e^{(I_{im} - I_{ij})}} \\ &= \frac{e^{I_{ij}}}{\sum_{m=1}^J e^{I_{im}}}. \end{aligned}$$

Again, we have the following conditions satisfied by the model: $\sum_{j=1}^J P_{ij} = 1$ and $0 < P_{ij} < 1$. The probability ratios (odds) are given by

$$\begin{aligned} \frac{P_{ij}}{P_{ir}} &= \frac{e^{I_{ij}}}{e^{I_{ir}}}, \text{ for } j \neq r \\ &= e^{(I_{ij} - I_{ir})}. \end{aligned}$$

In terms of log odds we have

$$\begin{aligned} \ln \left(\frac{P_{ij}}{P_{ir}} \right) &= I_{ij} - I_{ir} \\ &= (Z_{ij} - Z_{ir})' \beta \\ &= \sum_{k=1}^K \beta_k (Z_{ijk} - Z_{irk}) . \end{aligned}$$

Notice that there is no constant term β_0 in the conditional logit model because it would simply cancel out.

We next examine the marginal effects of the attributes on the choice probabilities. First consider the effect of the k th attribute (continuous) for choice j on the

probability of choosing j :

$$\frac{\partial P_{ij}}{\partial Z_{ijk}} = (P_{ij})(1 - P_{ij})\beta_k.$$

We can also express the cross-effect of the k th attribute for choice r on the probability of choosing j :

$$\frac{\partial P_{ij}}{\partial Z_{irk}} = -P_{ij}P_{ir}\beta_k$$

It should be clear that $\sum_{m=1}^J \left(\frac{\partial P_{ij}}{\partial Z_{imk}} \right) = 0$. If Z_{ijk} is a binary variable, then its marginal effect on P_{ij} would be calculated by

$$\Delta P_{ijk} = P_{ij(Z_{ijk}=1)} - P_{ij(Z_{ijk}=0)}.$$

The cross-effect of the k th binary attribute for choice r on the probability of choosing j is determined according to

$$\Delta P_{ijk_r} = P_{ij(Z_{irk}=1)} - P_{ij(Z_{irk}=0)}.$$

Estimation of the β 's are obtained by maximizing the likelihood function or log likelihood function for the sample:

$$L = \prod_{i=1}^n P_{i1}^{y_{i1}} \dots P_{iJ}^{y_{iJ}}$$

$$\ln(L) = \sum_{i=1}^n \sum_{j=1}^J y_{ij} \ln(P_{ij}),$$

where $y_{ij} = 0, 1$ is the actual binary outcome for the random variable Y_{ij} .

The estimated probabilities for each observation are obtained from the *MLE* estimated parameters:

$$\hat{P}_{ij} = \frac{e^{\hat{I}_{ij}}}{\sum_{m=1}^J e^{\hat{I}_{im}}},$$

where $\hat{I}_{ij} = \sum_{k=1}^K \hat{\beta}_k Z_{ijk}$ and $\hat{I}_{im} = \sum_{k=1}^K \hat{\beta}_k Z_{imk}$. Unlike the multinomial logit model, the sample average of the predicted probabilities from the conditional logit model will

not in general equal the sample proportions for the outcomes:

$$\begin{aligned}\bar{P}_j &= \frac{\sum_{i=1}^n \hat{P}_{ij}}{n} \\ &\neq \frac{n_j}{n} = \bar{Y}_j.\end{aligned}$$

How does one estimate the sample-wide marginal effects for the conditional logit model? One possibility is to obtain the estimated probabilities at the sample mean of the Z 's:

$$\hat{P}_j = \frac{e^{\bar{I}_j}}{\sum_{m=1}^J e^{\bar{I}_m}},$$

where $\bar{I}_j = \sum_{k=1}^K \hat{\beta}_k \bar{Z}_{jk}$, $\bar{I}_m = \sum_{k=1}^K \hat{\beta}_k \bar{Z}_{mk}$, $\bar{Z}_{jk} = \frac{\sum_{i=1}^n Z_{ijk}}{n}$, and $\bar{Z}_{mk} = \frac{\sum_{i=1}^n Z_{imk}}{n}$.

First consider the marginal effect of the k th attribute (continuous) for choice j on the probability of choosing j :

$$\frac{\partial \widehat{P}_j}{\partial Z_{jk}} = \left(\hat{P}_j \right) \left(1 - \hat{P}_j \right) \hat{\beta}_k.$$

for Z_{jk} a continuous variable. We can express the estimated cross-effect of the k th attribute for choice r on the probability of choosing j by:

$$\frac{\partial \widehat{P}_j}{\partial Z_{rk}} = -\hat{P}_j \hat{P}_r \hat{\beta}_k$$

In the case of a binary Z_{jk} variable, we have

$$\widehat{\Delta P}_{jk} = \hat{P}_{j(Z_{jk}=1)} - \hat{P}_{j(Z_{jk}=0)}.$$

The estimated cross-effect of the k th binary attribute for choice r on the probability of choosing j is determined according to

$$\widehat{\Delta P}_{jk_r} = \hat{P}_{j(Z_{rk}=1)} - \hat{P}_{j(Z_{rk}=0)}.$$

An equally valid alternative for estimating marginal effects is to calculate the sample average of all individual's marginal effects:

$$\frac{\partial P_j}{\partial Z_{jk}} = \frac{\sum_{i=1}^n \left(\frac{\partial P_{ij}}{\partial Z_{ijk}} \right)}{n}$$

for Z_{jk} a continuous variable and $\widehat{\frac{\partial P_{ij}}{\partial Z_{ijk}}} = \left(\hat{P}_{ij}\right) \left(1 - \hat{P}_{ij}\right) \hat{\beta}_k$. For estimated cross-effects of say the k th attribute for the r th choice on the probability of choosing j we could calculate

$$\overline{\frac{\partial P_j}{\partial Z_{rk}}} = \frac{\sum_{i=1}^n \left(\widehat{\frac{\partial P_{ij}}{\partial Z_{irk}}}\right)}{n},$$

where $\widehat{\frac{\partial P_{ij}}{\partial Z_{irk}}} = -\hat{P}_{ij}\hat{P}_{ir}\hat{\beta}_k$. If Z_{jk} is a binary variable, then its marginal effect on P_j could be calculated by

$$\overline{\Delta P_{jk}} = \frac{\sum_{i=1}^n \widehat{\Delta P_{ijk}}}{n}$$

where $\widehat{\Delta P_{ijk}} = \hat{P}_{ij(Z_{ijk}=1)} - \hat{P}_{ij(Z_{ijk}=0)}$. The estimated cross-effect of the k th binary attribute for choice r on the probability of choosing j is determined according to

$$\overline{\Delta P_{jk_r}} = \frac{\sum_{i=1}^n \widehat{\Delta P_{ijk_r}}}{n},$$

where $\widehat{\Delta P_{ijk_r}} = \hat{P}_{ij(Z_{irk}=1)} - \hat{P}_{ij(Z_{irk}=0)}$.

The pseudo R^2 and predictive success rates can be used to evaluate the goodness of fit of the conditional logit model. The conditional logit model can be used to predict the choice or outcome according to

$$\begin{aligned} \hat{Y}_{ij} &= 1 \text{ if } \hat{P}_{ij} = \max \left\{ \hat{P}_{i1}, \dots, \hat{P}_{iJ} \right\} \\ &= 0 \text{ otherwise.} \end{aligned}$$

The success rate = $\frac{\sum_{j=1}^J N_{jj}}{n}$, where N_{jj} = the number of actual j choices that were correctly predicted to be j choices.

C. Ordered Probit/Logit Model

When multiple outcomes or choices are ordered in some way, ordered probit/logit models are appropriate. Examples of ordered outcomes and choices include employee

assignments to job grade levels, preference rankings in opinion surveys (e.g. strongly agree, mildly agree, neither agree or disagree, mildly disagree, strongly disagree) etc.

Ordered outcomes can be motivated by a latent variable specification:

$$Y_i^* = I_i + \varepsilon_i, \quad i = 1, \dots, n$$

where $I_i = X_i' \beta = \beta_0 + \sum_{k=1}^K \beta_k X_{ik}$ is the index function and ε_i is an independently and identically distributed error terms satisfying all of the classical assumptions. Y_i^* is a latent variable, examples of which might include the utility of an outcome, intensity of opinions, value of qualifications etc. Although Y_i^* is not directly observed, a set of binary outcome variables Y_{ij} related to Y_i^* are observed. These binary variables correspond to $J + 1$ observed choice or outcome categories:

$$\begin{aligned} Y_{i0} &= 1 [Y_i^* \leq 0] \\ Y_{i1} &= 1 [0 < Y_i^* \leq \mu_1] \\ Y_{i2} &= 1 [\mu_1 < Y_i^* \leq \mu_2] \\ &\cdot \\ &\cdot \\ &\cdot \\ Y_{iJ} &= 1 [\mu_{J-1} < Y_i^*] \end{aligned}$$

where μ_j is a threshold parameter, and Y_{ij} is an indicator variable that takes on the value of 1 if the corresponding indicator condition is met, e.g. $Y_{i2} = 1$ if $\mu_1 < Y_i^* \leq \mu_2$ and = 0 otherwise, which is compactly expressed as $Y_{i2} = 1 [\mu_1 < Y_i^* \leq \mu_2]$. The threshold parameters are ordered according to $0 < \mu_1 < \mu_2 < \dots < \mu_{J-1}$.

An equivalent way to define the ordered outcomes is to define the observed variable

Y_i according to

$$\begin{aligned}
Y_i &= 0 \text{ if } Y_i^* \leq 0 \\
&= 1 \text{ if } 0 < Y_i^* \leq \mu_1 \\
&= 2 \text{ if } \mu_1 < Y_i^* \leq \mu_2 \\
&\cdot \\
&\cdot \\
&\cdot \\
&= J \text{ if } \mu_{J-1} < Y_i^*.
\end{aligned}$$

Variable Y_i can therefore assume $J + 1$ distinct values.

The probabilities associated with each outcome for each individual in the sample are specified according to

$$\begin{aligned}
\text{prob}(Y_{i0} = 1) &= \text{prob}(Y_i^* \leq 0) = \text{prob}(\varepsilon_i \leq -I_i) = F(-I_i) \\
\text{prob}(Y_{i1} = 1) &= \text{prob}(0 < Y_i^* \leq \mu_1) = \text{prob}(-I_i < \varepsilon_i \leq \mu_1 - I_i) \\
&= F(\mu_1 - I_i) - F(-I_i) \\
\text{prob}(Y_{i2} = 1) &= \text{prob}(\mu_1 < Y_i^* \leq \mu_2) = \text{prob}(\mu_1 - I_i < \varepsilon_i \leq \mu_2 - I_i) \\
&= F(\mu_2 - I_i) - F(\mu_1 - I_i) \\
&\cdot \\
&\cdot \\
&\cdot \\
\text{prob}(Y_{iJ} = 1) &= \text{prob}(\mu_{J-1} < Y_i^*) = \text{prob}(\mu_{J-1} - I_i < \varepsilon_i) = 1 - F(\mu_{J-1} - I_i),
\end{aligned}$$

where $F(\cdot)$ is the cdf which could either be $\Phi(\cdot)$, the standard normal, or $\Lambda(\cdot)$, the logistic.

As in the cases with the probit and logit models, the β parameters are not the marginal effects of the variables on the probabilities. First consider the marginal

effects of continuous variable X_{ik} :

$$\begin{aligned}
\frac{\partial P_{i0}}{\partial X_{ik}} &= -f(I_i) \beta_k \\
\frac{\partial P_{i1}}{\partial X_{ik}} &= [f(I_i) - f(\mu_1 - I_i)] \beta_k \\
\frac{\partial P_{i2}}{\partial X_{ik}} &= [f(\mu_1 - I_i) - f(\mu_2 - I_i)] \beta_k \\
&\cdot \\
&\cdot \\
&\cdot \\
\frac{\partial P_{iJ}}{\partial X_{ik}} &= [f(\mu_{J-1} - I_i)] \beta_k,
\end{aligned}$$

where $P_{ij} = \text{prob}(Y_{ij} = 1)$. Note that apart from $\frac{\partial P_{i0}}{\partial X_{ik}}$ and $\frac{\partial P_{iJ}}{\partial X_{ik}}$, one cannot determine the sign of the marginal effect from the sign of the coefficient β_k alone. If X_{ik} is a discrete (binary) variable, the marginal effects are given by

$$\Delta P_{ij_k} = P_{ij}(X_{ik}=1) - P_{ij}(X_{ik}=0).$$

The β and μ parameters are estimated by MLE from the likelihood function or log likelihood function for the sample:

$$\begin{aligned}
L &= \prod_{i=1}^n \{ [F(-I_i)]^{y_{i0}} [F(\mu_1 - I_i) - F(-I_i)]^{y_{i1}} [F(\mu_2 - I_i) - F(\mu_1 - I_i)]^{y_{i2}} \\
&\quad \cdots [1 - F(\mu_{J-1} - I_i)]^{y_{iJ}} \}, \\
\ln(L) &= \sum_{i=1}^n \{ y_{i0} \ln [F(-I_i)] + y_{i1} \ln [F(\mu_1 - I_i) - F(-I_i)] \\
&\quad + y_{i2} \ln [F(\mu_2 - I_i) - F(\mu_1 - I_i)] \\
&\quad + \cdots + y_{iJ} \ln [1 - F(\mu_{J-1} - I_i)] \}.
\end{aligned}$$

In the STATA econometrics program instead of reporting the estimated μ_j parameters

directly, the program reports

$$\begin{aligned}\text{cut 1} &= -\hat{\beta}_0 \\ \text{cut 2} &= \hat{\mu}_1 - \hat{\beta}_0 \\ &\cdot \\ &\cdot \\ &\cdot \\ \text{cut J-1} &= \hat{\mu}_{J-1} - \hat{\beta}_0.\end{aligned}$$

It is straightforward to solve for the values of $\hat{\mu}_j$:

$$\begin{aligned}\hat{\beta}_0 &= -\text{cut 1} \\ \hat{\mu}_1 &= \text{cut 2} + \hat{\beta}_0 \\ &\cdot \\ &\cdot \\ &\cdot \\ \hat{\mu}_{J-1} &= \text{cut J-1} + \hat{\beta}_0.\end{aligned}$$

As before, the marginal effects for the sample can be estimated at the sample mean values of the variables or as the sample average of the estimated individual marginal effects. Evaluating at the sample mean yields the following probability estimates:

$$\begin{aligned}
\hat{P}_0 &= F(-\bar{I}) \\
\hat{P}_1 &= F(\hat{\mu}_1 - \bar{I}) - F(-\bar{I}) \\
\hat{P}_2 &= F(\hat{\mu}_2 - \bar{I}) - F(\hat{\mu}_1 - \bar{I}) \\
&\cdot \\
&\cdot \\
&\cdot \\
\hat{P}_J &= 1 - F(\hat{\mu}_{J-1} - \bar{I}),
\end{aligned}$$

where $\bar{I} = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k \bar{X}_k$. The corresponding estimated marginal effects for a continuous variable X_k are calculated according to

$$\begin{aligned}
\frac{\widehat{\partial P_0}}{\partial X_k} &= -f(\bar{I}) \hat{\beta}_k \\
\frac{\widehat{\partial P_1}}{\partial X_k} &= [f(\bar{I}) - f(\hat{\mu}_1 - \bar{I})] \hat{\beta}_k \\
\frac{\widehat{\partial P_2}}{\partial X_k} &= [f(\hat{\mu}_1 - \bar{I}) - f(\hat{\mu}_2 - \bar{I})] \hat{\beta}_k \\
&\cdot \\
&\cdot \\
&\cdot \\
\frac{\widehat{\partial P_J}}{\partial X_k} &= [f(\hat{\mu}_{J-1} - \bar{I})] \hat{\beta}_k,
\end{aligned}$$

In the case of a binary variable X_k , the marginal effects on each probability are calculated according to $\widehat{\Delta P}_{j-k} = \hat{P}_{j(X_k=1)} - \hat{P}_{j(X_k=0)}$.

The alternative method for calculating sample probabilities and marginal effects

are to simply average the estimates over the sample:

$$\begin{aligned}
\overline{P_0} &= \frac{\sum_{i=1}^n F(-\hat{I}_i)}{n} \\
\overline{P_1} &= \frac{\sum_{i=1}^n [F(\hat{\mu}_1 - \hat{I}_i) - F(-\hat{I}_i)]}{n} \\
\overline{P_2} &= \frac{\sum_{i=1}^n [F(\hat{\mu}_2 - \hat{I}_i) - F(\hat{\mu}_1 - \hat{I}_i)]}{n} \\
&\cdot \\
&\cdot \\
&\cdot \\
\overline{P_J} &= \frac{\sum_{i=1}^n [1 - F(\hat{\mu}_{J-1} - \hat{I}_i)]}{n},
\end{aligned}$$

where $\hat{I}_i = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k X_{ik}$. The sample averages for the marginal effects are given by

$$\begin{aligned}
\frac{\widehat{\partial P_0}}{\partial X_k} &= \frac{\sum_{i=1}^n \frac{\partial \widehat{P_{i0}}}{\partial X_{ik}}}{n} \\
\frac{\widehat{\partial P_1}}{\partial X_k} &= \frac{\sum_{i=1}^n \frac{\partial \widehat{P_{i1}}}{\partial X_{ik}}}{n} \\
\frac{\widehat{\partial P_2}}{\partial X_k} &= \frac{\sum_{i=1}^n \frac{\partial \widehat{P_{i2}}}{\partial X_{ik}}}{n} \\
&\cdot \\
&\cdot \\
&\cdot \\
\frac{\widehat{\partial P_J}}{\partial X_k} &= \frac{\sum_{i=1}^n \frac{\partial \widehat{P_{iJ}}}{\partial X_{ik}}}{n},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \widehat{P}_{i0}}{\partial X_{ik}} &= -f(\hat{I}_i) \hat{\beta}_k \\
\frac{\partial \widehat{P}_{i1}}{\partial X_{ik}} &= [f(I_i) - f(\hat{\mu}_1 - I_i)] \hat{\beta}_k \\
\frac{\partial \widehat{P}_{i2}}{\partial X_{ik}} &= [f(\hat{\mu}_1 - I_i) - f(\hat{\mu}_2 - I_i)] \hat{\beta}_k \\
&\cdot \\
&\cdot \\
&\cdot \\
\frac{\partial \widehat{P}_{iJ}}{\partial X_{ik}} &= [f(\hat{\mu}_{J-1} - I_i)] \hat{\beta}_k.
\end{aligned}$$

For a binary variable X_{ik} , the marginal effect on each probability would be calculated

$$\text{as } \overline{\Delta P_{j_k}} = \frac{\sum_{i=1}^n [\widehat{P}_{ij(X_{ik}=1)} - \widehat{P}_{ij(X_{ik}=0)}]}{n}.$$

IV. LIMITED DEPENDENT VARIABLE MODELS

Background Reading:

Greene W.H., *Econometric Analysis*, 6th ed., Chapter 24 (pp.863-881, 882-889).

Wooldridge, J.M., *Introductory Econometrics: A Modern Approach*, 4th ed., Chapter 17 (pp. 587-595, 600-613).

Heckman, J., "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, vol. 47, pp.153-161.

Tobin, J., "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 1958, pp.24-36.

A. Truncated Regression Model

Suppose we have a conventional linear model defined by

$$Y_i = I_i + \varepsilon_i$$

where $I_i = X_i' \beta = \beta_0 + \sum_{k=1}^K \beta_k X_{ik}$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and satisfies all of the classical assumptions. However, suppose our data set pertains only to observations for which $Y_i > a$. An example might be where Y refers to income and $a = \$50,000$. Another example might be where Y refers to hours worked during the survey year and $a = 1,000$ hours. In the general population

$$\begin{aligned} \text{prob}(Y_i > a) &= \text{prob}(I_i + \varepsilon_i > a) \\ &= \text{prob}(\varepsilon_i > a - I_i) \\ &= 1 - \text{prob}(\varepsilon_i < a - I_i). \end{aligned}$$

For the subpopulation $Y_i > a$ contained in our truncated data set, the truncated error term $\varepsilon_i | \varepsilon_i > a - I_i$ follows a truncated normal distribution with mean $\sigma_\varepsilon \lambda_i > 0$

and variance $\sigma_\varepsilon^2 [1 - \delta(z_{ia})] < \sigma_\varepsilon^2$, where $\lambda_i = \frac{\phi(z_{ia})}{1 - \Phi(z_{ia})}$, $z_{ia} = \frac{a - I_i}{\sigma_\varepsilon}$, and $\delta(z_{ia}) = (\lambda_i)(\lambda_i - z_{ia})$.

The actual regression model for our truncated sample becomes

$$Y_i | Y_i > a = I_i + \varepsilon_i | (\varepsilon_i > a - I_i).$$

The classical assumptions do not hold for this regression model because even though ε_i is uncorrelated with the X 's in the original model, the truncated error term $\varepsilon_i | (\varepsilon_i > a - I_i)$ is not independent of the X 's and $E(\varepsilon_i | \varepsilon_i > a - I_i) = \sigma_\varepsilon \lambda_i \neq 0$. Furthermore, the variance of $\varepsilon_i | (\varepsilon_i > a - I_i)$ is not constant, i.e. $var(\varepsilon_i | \varepsilon_i > a - I_i) = \sigma_\varepsilon^2 [1 - \delta(z_{ia})] < \sigma_\varepsilon^2$ even though the variance of ε_i in the original model is constant, i.e. $var(\varepsilon_i) = \sigma_\varepsilon^2$. This is the problem of heteroscedasticity. What this all means is that ordinary least squares (*OLS*) would not be a suitable estimator for the model. The *OLS* estimator would be biased and inconsistent.

While the β 's are the marginal effects on Y_i from the original population, these parameters are not the marginal effects on the Y 's we observe in our data set's truncated sample. We can see that the expected value of our truncated dependent variable $Y_i | Y_i > a$ (conditioning on the X 's in I_i) is given by

$$\begin{aligned} E(Y_i | Y_i > a) &= I_i + E(\varepsilon_i | \varepsilon_i > a - I_i) \\ &= I_i + \sigma_\varepsilon \lambda_i. \end{aligned}$$

The marginal effect of a continuous variable X_{ik} on $E(Y_i | Y_i > a)$ turns out (after some calculus and algebra) to be

$$\frac{\partial E(Y_i | Y_i > a)}{\partial X_{ik}} = [1 - \delta(z_{ia})] \beta_k.$$

In absolute value $|[1 - \delta(z_{ia})] \beta_k| < |\beta_k|$ which makes sense because the truncated sample reduces the variation in the observed Y 's. For a binary variable X_{ik} , the marginal effect is given by

$$\Delta [E(Y_i | Y_i > a)]_k = E(Y_i | Y_i > a)_{(X_{ik}=1)} - E(Y_i | Y_i > a)_{(X_{ik}=0)}.$$

Consistent estimators of the model are obtained by maximizing the likelihood function for the truncated sample. The likelihood function and log likelihood function for the sample are given by

$$L = \prod_{i=1}^n \frac{\left(\frac{1}{\sigma_\varepsilon}\right) \phi \left[\left(\frac{Y_i - I_i}{\sigma_\varepsilon}\right) \right]}{1 - \Phi(z_{ia})}.$$

and

$$\ln(L) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n (Y_i - I_i)^2 - \sum_{i=1}^n \ln[1 - \delta(z_{ia})].$$

As usual, the estimated marginal effects for the sample can be obtained in two different ways. If we evaluate everything at the sample mean, the estimated conditional mean value of the truncated dependent variable is given by

$$\begin{aligned} E(\widehat{Y|Y > a}) &= \bar{I} + \hat{\sigma}_\varepsilon \hat{\lambda} \\ &\neq \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}, \end{aligned}$$

where $\bar{I} = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k \bar{X}_k$, $\hat{\lambda} = \frac{\phi(\hat{z}_a)}{1 - \Phi(\hat{z}_a)}$, and $\hat{z}_a = \frac{a - \bar{I}}{\hat{\sigma}_\varepsilon}$. The marginal effects for a continuous variable X_k are calculated according to

$$\frac{\partial E(\widehat{Y|Y > a})}{\partial X_k} = [1 - \delta(\hat{z}_a)] \hat{\beta}_k,$$

where $\delta(\hat{z}_a) = \left(\hat{\lambda}\right) \left(\hat{\lambda} - \hat{z}_a\right)$. The estimated marginal effect for a binary variable X_k would be calculated as

$$\Delta [E(\widehat{Y|Y > a})]_k = E(Y|\widehat{Y > a})_{(X_k=1)} - E(Y|\widehat{Y > a})_{(X_k=0)}$$

where $E(Y|\widehat{Y > a})_{(X_k=1)} = \left[\bar{I}_{(X_k=1)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{(X_k=1)}\right]$ and $E(Y|\widehat{Y > a})_{(X_k=0)} = \left[\bar{I}_{(X_k=0)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{(X_k=0)}\right]$.

The estimated truncated mean and marginal effects can alternatively be calculated as the sample mean of the estimated marginal effects across the sample observations:

$$\overline{E(Y|Y > a)} = \frac{\sum_{i=1}^n E(Y_i|Y_i > a)}{n}$$

where $E(\widehat{Y_i|Y_i > a}) = \hat{I}_i + \hat{\sigma}_\varepsilon \hat{\lambda}_i$, $\hat{I}_i = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k X_{ik}$, $\hat{\lambda}_i = \frac{\phi(\hat{z}_{ia})}{1 - \Phi(\hat{z}_{ia})}$, and $\hat{z}_{ia} = \frac{a - \hat{I}_i}{\hat{\sigma}_\varepsilon}$. The estimated marginal effects are calculated according to

$$\frac{\partial E(Y|Y > a)}{\partial X_k} = \frac{\sum_{i=1}^n \frac{\partial E(\widehat{Y_i|Y_i > a})}{\partial X_{ik}}}{n},$$

where $\frac{\partial E(\widehat{Y_i|Y_i > a})}{\partial X_{ik}} = [1 - \delta(\hat{z}_{ia})] \hat{\beta}_k$, and $\delta(\hat{z}_{ia}) = \left(\hat{\lambda}_i\right) \left(\hat{\lambda}_i - \hat{z}_{ia}\right)$. The estimated marginal effect for a binary variable X_k would be calculated as

$$\Delta [E(Y|Y > a)]_k = \frac{\sum_{i=1}^n \Delta [E(\widehat{Y_i|Y_i > a})]_k}{n},$$

where

$$\begin{aligned} \Delta [E(\widehat{Y_i|Y_i > a})]_k &= E(Y_i|\widehat{Y_i > a})_{(X_{ik}=1)} - E(Y_i|\widehat{Y_i > a})_{(X_{ik}=0)} \\ &= \left[\hat{I}_{i(X_{ik}=1)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{i(X_{ik}=1)} \right] - \left[\hat{I}_{i(X_{ik}=0)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{i(X_{ik}=0)} \right]. \end{aligned}$$

The analysis is very similar when there is truncation from above, i.e. our data set consists only of observations for which $Y_i < a$. In this case $\lambda_i = -\frac{\phi(z_{ia})}{\Phi(z_{ia})}$. The truncated regression model for this sample is written as

$$Y_i|Y_i < a = I_i + \varepsilon_i | (\varepsilon_i < a - I_i),$$

where $E(\varepsilon_i | \varepsilon_i < a - I_i) = \sigma_\varepsilon \lambda_i < 0$.

B. Tobit (Censored) Regression Model

Suppose we have a seemingly conventional linear model defined by

$$Y_i^* = I_i + \varepsilon_i, \quad i = 1, \dots, n$$

where $I_i = X_i' \beta = \beta_0 + \sum_{k=1}^K \beta_k X_{ik}$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and satisfies all of the classical assumptions. However, the variable Y_i^* is not observed for the entire sample. Instead we observe the variable Y_i defined by

$$\begin{aligned} Y_i &= 0 \text{ if } Y_i^* \leq 0 \text{ (limit observations)} \\ &= Y_i^* \text{ if } Y_i^* > 0 \text{ (nonlimit observations)}. \end{aligned}$$

The regression model becomes

$$\begin{aligned} Y_i &= I_i + \varepsilon_i \text{ if } I_i + \varepsilon_i > 0 \text{ (or } \varepsilon_i > -I_i) \\ &= 0 \text{ otherwise (} \varepsilon_i \leq -I_i). \end{aligned}$$

Examples for the use of a Tobit model include modeling the determinants of prison sentences in which some convicted individuals receive 0 prison time. Another example is the Tobin (1958) case in which one wishes to model the determinants of how much individuals would pay for a new car. The sample includes individuals who did not purchase a new car so the amount paid is coded as 0. In these cases, *OLS* estimation of the model would be biased and inconsistent even if the limit observations were dropped from the sample. In this latter case, we would have the truncated regression model in which $a = 0$. So the censored or Tobit model uses more information since we at least observe the X_i 's for the limit observations.

Depending on what one is interested in, there are a variety of conditional mean values for the dependent variable that one could consider:

$$\begin{aligned} E(Y_i^* | I_i) &= I_i \\ &= \beta_0 + \sum_{k=1}^K \beta_k X_{ik} \\ &\text{(Population Regression Function)} \end{aligned}$$

$$E(Y_i | I_i, Y_i = 0) = 0$$

$$\begin{aligned}
E(Y_i|I_i, Y_i > 0) &= I_i + E(\varepsilon_i|\varepsilon_i > -I_i) \\
&= I_i + \sigma_\varepsilon \lambda_i,
\end{aligned}$$

where $\lambda_i = \frac{\phi\left(\frac{-I_i}{\sigma_\varepsilon}\right)}{1 - \Phi\left(\frac{-I_i}{\sigma_\varepsilon}\right)} = \frac{\phi\left(\frac{I_i}{\sigma_\varepsilon}\right)}{\Phi\left(\frac{I_i}{\sigma_\varepsilon}\right)}$. The expected value of Y_i given only I_i can be expressed as a weighted average of $E(Y_i|I_i, Y_i = 0)$ and $E(Y_i|I_i, Y_i > 0)$ with the weights equal to the probabilities of $Y_i = 0$ and $Y_i > 0$:

$$\begin{aligned}
E(Y_i|I_i) &= E(Y_i|I_i, Y_i = 0) \cdot \Phi\left(\frac{-I_i}{\sigma_\varepsilon}\right) + E(Y_i|I_i, Y_i > 0) \cdot \left[1 - \Phi\left(\frac{-I_i}{\sigma_\varepsilon}\right)\right] \\
&= E(Y_i|I_i, Y_i > 0) \cdot \Phi\left(\frac{I_i}{\sigma_\varepsilon}\right) \\
&= (I_i + \sigma_\varepsilon \lambda_i) \cdot \Phi\left(\frac{I_i}{\sigma_\varepsilon}\right),
\end{aligned}$$

where $\left[1 - \Phi\left(\frac{-I_i}{\sigma_\varepsilon}\right)\right] = \Phi\left(\frac{I_i}{\sigma_\varepsilon}\right)$ is the probability of drawing a nonlimit observation, i.e. $\text{prob}(Y_i > 0)$.

The marginal effects of the X 's on the conditional means also vary, depending upon what one is interested in. For the population the marginal effects are simply

$$\frac{\partial E(Y_i^*|I_i)}{\partial X_{ik}} = \beta_k$$

for X_{ik} continuous or binary. For the nonlimit observations, the marginal effects are the same as for the truncated regression model:

$$\frac{\partial E(Y_i|I_i, Y_i > 0)}{\partial X_{ik}} = \left[1 - \lambda_i \left(\frac{I_i}{\sigma_\varepsilon} + \lambda_i\right)\right] \beta_k$$

for X_{ik} a continuous variable and

$$\Delta [E(Y_i|Y_i > 0)]_k = E(Y_i|Y_i > 0)_{(X_{ik}=1)} - E(Y_i|Y_i > 0)_{(X_{ik}=0)},$$

for X_{ik} a binary variable. The marginal effects on Y_i for the censored sample are given by

$$\frac{\partial E(Y_i|I_i)}{\partial X_{ik}} = \Phi\left(\frac{I_i}{\sigma_\varepsilon}\right) \beta_k$$

for X_{ik} a continuous variable, and

$$\begin{aligned}\Delta [E(Y_i|I_i)]_k &= [E(Y_i|I_i(X_{ik}=1))]_k - [E(Y_i|I_i(X_{ik}=0))]_k \\ &= \left[(I_i(X_{ik}=1) + \sigma_\varepsilon \lambda_{i(X_{ik}=1)}) \cdot \Phi\left(\frac{I_i(X_{ik}=1)}{\sigma_\varepsilon}\right) \right] \\ &\quad - \left[(I_i(X_{ik}=0) + \sigma_\varepsilon \lambda_{i(X_{ik}=0)}) \cdot \Phi\left(\frac{I_i(X_{ik}=0)}{\sigma_\varepsilon}\right) \right],\end{aligned}$$

for X_{ik} a binary variable.

Consistent estimators of the model are obtained by maximizing the likelihood function for the censored sample. The likelihood function and log likelihood function for the sample are given by

$$L = \prod_{y_i=0} \left[1 - \Phi\left(\frac{I_i}{\sigma_\varepsilon}\right) \right] \prod_{y_i>0} \left(\frac{1}{\sigma_\varepsilon} \right) \phi \left[\left(\frac{Y_i - I_i}{\sigma_\varepsilon} \right) \right]$$

and

$$\ln(L) = \sum_{y_i=0} \ln \left[1 - \Phi\left(\frac{I_i}{\sigma_\varepsilon}\right) \right] - \frac{1}{2} \sum_{y_i>0} \left[\ln(2\pi) + \ln(\sigma_\varepsilon^2) + \frac{(Y_i - I_i)^2}{\sigma_\varepsilon^2} \right].$$

In the Tobit model, the likelihood function is a mixture of probabilities (for the limit observations) and density functions (for the nonlimit observations).

Again the estimated marginal effects for the sample can be obtained in two different ways. If we evaluate everything at the sample mean, the estimated conditional population mean can be calculated as

$$E(\widehat{Y^*|I}) = \bar{I}$$

where $\bar{I} = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k \bar{X}_k$ and $\bar{X}_k = \frac{\sum_{i=1}^n X_{ik}}{n}$. The estimated truncated mean can be calculated as:

$$E(Y|\widehat{I}, Y > 0) = \bar{I} + \hat{\sigma}_\varepsilon \hat{\lambda},$$

where $\hat{\lambda} = \frac{\phi\left(\frac{\bar{I}}{\hat{\sigma}_\varepsilon}\right)}{\Phi\left(\frac{\bar{I}}{\hat{\sigma}_\varepsilon}\right)}$. The estimated censored conditional mean is calculated according to

$$E(\widehat{Y|I}) = \left(\bar{I} + \hat{\sigma}_\varepsilon \hat{\lambda}\right) \cdot \Phi\left(\frac{\bar{I}}{\hat{\sigma}_\varepsilon}\right)$$

Marginal effects evaluated at the overall sample mean characteristics are calculated according to

$$\frac{\partial E(Y^*|I)}{\partial X_k} = \hat{\beta}_k,$$

for any variable X_k ,

$$\frac{\partial E(\widehat{Y|I, Y > 0})}{\partial X_k} = \left[1 - \hat{\lambda} \left(\frac{\bar{I}}{\hat{\sigma}_\varepsilon} + \hat{\lambda}\right)\right] \hat{\beta}_k$$

for the truncated mean with continuous variable X_k , and

$$\Delta [E(\widehat{Y|Y > 0})]_k = E(Y|\widehat{Y > 0})_{(X_{ik}=1)} - E(Y|\widehat{Y > 0})_{(X_{ik}=0)}$$

for the binary variable X_k , where

$$E(Y|\widehat{Y > 0})_{(X_k=1)} = \bar{I}_{(X_k=1)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{(X_k=1)}$$

$$E(Y|\widehat{Y > 0})_{(X_k=0)} = \bar{I}_{(X_k=0)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{(X_k=0)},$$

and for the censored mean

$$\frac{\partial E(\widehat{Y|I})}{\partial X_k} = \Phi\left(\frac{\bar{I}}{\hat{\sigma}_\varepsilon}\right) \hat{\beta}_k$$

for continuous variable X_k , and

$$\Delta [E(\widehat{Y|I})]_k = [E(\widehat{Y|I}_{(X_k=1)})]_k - [E(\widehat{Y|I}_{(X_k=0)})]_k$$

for a binary variable X_k , where $[E(\widehat{Y|I}_{(X_k=1)})]_k = \left[\left(\bar{I}_{(X_k=1)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{(X_k=1)}\right) \cdot \Phi\left(\frac{\bar{I}_{(X_k=1)}}{\hat{\sigma}_\varepsilon}\right)\right]$

and $[E(\widehat{Y|I}_{(X_k=0)})]_k = \left[\left(\bar{I}_{(X_k=0)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{(X_k=0)}\right) \cdot \Phi\left(\frac{\bar{I}_{(X_k=0)}}{\hat{\sigma}_\varepsilon}\right)\right]$.

Our second method for obtaining estimated conditional means and marginal effects is to use sample averages of individual predicted means and marginal effects. The estimated conditional population mean is the same as evaluation at the sample mean, i.e. $\widehat{E}(Y^*|I) = \bar{I}$. The estimated truncated mean in this case is calculated as:

$$\overline{E(Y|I, Y > 0)} = \bar{I} + \hat{\sigma}_\varepsilon \bar{\lambda},$$

where $\bar{\lambda} = \frac{\sum_{i=1}^n \hat{\lambda}_i}{n}$, $\hat{\lambda}_i = \frac{\phi\left(\frac{\hat{I}_i}{\hat{\sigma}_\varepsilon}\right)}{\Phi\left(\frac{\hat{I}_i}{\hat{\sigma}_\varepsilon}\right)}$, and $\hat{I}_i = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k X_{ik}$. For the estimated censored sample mean we have

$$\overline{E(Y|I)} = \frac{\sum_{i=1}^n E(\widehat{Y}_i|I_i)}{n},$$

where $E(\widehat{Y}_i|I_i) = \left(\hat{I}_i + \hat{\sigma}_\varepsilon \hat{\lambda}_i\right) \cdot \Phi\left(\frac{\hat{I}_i}{\hat{\sigma}_\varepsilon}\right)$.

Estimated marginal effects for the population mean are the same as when using the sample mean characteristics, i.e. $\frac{\partial E(Y^*|I)}{\partial X_k} = \hat{\beta}_k$ for any variable X_k . Marginal effects for the truncated mean are calculated for a continuous variable X_k according to

$$\frac{\partial E(\widehat{Y}|I, Y > 0)}{\partial X_k} = \frac{\sum_{i=1}^n \frac{\partial E(\widehat{Y}_i|I_i, Y_i > 0)}{\partial X_{ik}}}{n},$$

where

$$\frac{\partial E(\widehat{Y}_i|I_i, Y_i > 0)}{\partial X_{ik}} = \left[1 - \hat{\lambda}_i \left(\frac{\hat{I}_i}{\hat{\sigma}_\varepsilon} + \hat{\lambda}_i \right) \right] \hat{\beta}_k,$$

and for a binary variable X_k according to

$$\Delta [E(\widehat{Y}|Y > 0)]_k = E(Y|\widehat{Y} > 0)_{(X_k=1)} - E(Y|\widehat{Y} > 0)_{(X_k=0)}$$

where

$$E(Y|\widehat{Y} > 0)_{(X_k=1)} = \frac{\sum_{i=1}^n E(\widehat{Y}_i|Y_i > 0)_{(X_{ik}=1)}}{n},$$

$$E(Y_i | \widehat{Y_i > 0})_{(X_{ik}=1)} = \hat{I}_{i(X_{ik}=1)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{i(X_{ik}=1)},$$

$$E(Y | \widehat{Y > 0})_{(X_k=0)} = \frac{\sum_{i=1}^n E(Y_i | \widehat{Y_i > 0})_{(X_{ik}=0)}}{n},$$

and

$$E(Y_i | \widehat{Y_i > 0})_{(X_{ik}=0)} = \hat{I}_{i(X_{ik}=0)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{i(X_{ik}=0)}.$$

Marginal effects for the censored sample mean are calculated according to

$$\frac{\partial \widehat{E(Y|I)}}{\partial X_k} = \frac{\sum_{i=1}^n \frac{\partial E(Y_i | I_i)}{\partial X_{ik}}}{n}$$

for X_k a continuous variable, where

$$\frac{\partial E(Y_i | I_i)}{\partial X_{ik}} = \Phi\left(\frac{\hat{I}_i}{\hat{\sigma}_\varepsilon}\right) \hat{\beta}_k,$$

and by

$$\Delta [E(\widehat{Y|I})]_k = [E(Y | \widehat{I}_{(X_k=1)})]_k - [E(Y | \widehat{I}_{(X_k=0)})]_k$$

for a binary variable X_k , where

$$[E(\widehat{Y|I}_{(X_k=1)})]_k = \frac{\sum_{i=1}^n [E(Y_i | \widehat{I}_{i(X_{ik}=1)})]_k}{n},$$

$$[E(Y_i | \widehat{I}_{i(X_{ik}=1)})]_k = \left[\left(\hat{I}_{i(X_{ik}=1)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{i(X_{ik}=1)} \right) \cdot \Phi\left(\frac{\hat{I}_{i(X_{ik}=1)}}{\hat{\sigma}_\varepsilon}\right) \right],$$

$$[E(\widehat{Y|I}_{(X_k=0)})]_k = \frac{\sum_{i=1}^n [E(Y_i | \widehat{I}_{i(X_{ik}=0)})]_k}{n},$$

and

$$[E(Y_i | \widehat{I}_{i(X_{ik}=0)})]_k = \left[\left(\hat{I}_{i(X_{ik}=0)} + \hat{\sigma}_\varepsilon \hat{\lambda}_{i(X_{ik}=0)} \right) \cdot \Phi\left(\frac{\hat{I}_{i(X_{ik}=0)}}{\hat{\sigma}_\varepsilon}\right) \right].$$

C. Heckit (Sample Selection) Regression Model

Sample selection models account for the nonrandomness often found in regression samples. The nonrandomness of the sample is addressed by accounting for the probability of an observation being nonrandomly selected into the regression sample. The full sample includes observations for both the selected regression sample and the non-selected sample. The Heckit procedure estimates the model in two steps. The first step consists of estimating a probit model for selection into the regression sample. The second step consists of estimating the regression with the Inverse Mills Ratio (IMR) added as an extra regressor to account for the probability of selection. The Heckit method differs from Tobit in two ways. First, the set of explanatory variables in the probit stage are generally not identical to the list of variables in the regression. Second, the censoring threshold is not a constant but varies with the probit variables that determine whether or not the dependent variable will be observed. Examples of sample selection models include estimating earnings equations for individuals observed to be in the labor force (employed) after taking account that employment and labor force participation are not random. Another example is trying to estimate the effects of migration on earnings. One would first estimate the probit model for the migration decision and then calculate the appropriate IMR for a migrant or a nonmigrant to add to the earnings equation.

Consider the following model:

$$\begin{aligned} H_i^* &= W_i + u_i \\ &= Z_i' \gamma + u_i \end{aligned}$$

$$\begin{aligned} Y_i &= I_i + \varepsilon_i \\ &= X_i' \beta + \varepsilon_i \end{aligned}$$

where $Z_i'\gamma = \gamma_0 + \sum_{j=1}^J \gamma_j Z_{ij}$, $X_i'\beta = \beta_0 + \sum_{k=1}^K \beta_k X_{ik}$, and (u_i, ε_i) follows a bivariate normal distribution with 0 means, correlation ρ , and variances σ_u^2 and σ_ε^2 .

Suppose H_i^* is a latent variable, e.g. the net benefit from working. However, suppose we observe the indicator variable H_i :

$$\begin{aligned} H_i &= 1 \text{ if } W_i + u_i > 0 \text{ (or } u_i > -W_i) \\ &= 0 \text{ if } W_i + u_i \leq 0 \text{ (or } u_i \leq -W_i). \end{aligned}$$

Since σ_u is not identified, we normalize the standard deviation of u_i by setting $\sigma_u = 1$. Because of this normalization, the covariance between u_i and ε_i is simply $\sigma_{\varepsilon u} = \sigma_{u\varepsilon} = \rho\sigma_\varepsilon\sigma_u = \rho\sigma_\varepsilon$. The sample selection rule is that Y_i is observed only when $H_i = 1$. It is clear that $\text{prob}(H_i = 1|W_i) = \Phi(W_i)$ is the probability that Y_i will be observed in the sample and $\text{prob}(H_i = 0|W_i) = 1 - \Phi(W_i)$ is the probability that Y_i will not be observed in the sample.

The regression model for the selected subsample for which Y_i is observed may be written as

$$Y_i | (H_i = 1) = I_i + \varepsilon_i | (H_i = 1), \quad i = 1, \dots, n_1,$$

where n_1 is the sample size for the selected observations. The expected value of ε_i in the selected sample is not equal to 0 even though its population mean is 0:

$$\begin{aligned} E(\varepsilon_i | H_i = 1) &= E(\varepsilon_i | u_i > -W_i) \\ &= \rho\sigma_\varepsilon\lambda_i \neq 0, \end{aligned}$$

where $\lambda_i = \frac{\phi(-W_i)}{1 - \Phi(-W_i)} = \frac{\phi(W_i)}{\Phi(W_i)}$ is the IMR. Note that the variance of ε_i in the selected sample is less than the population variance σ_ε^2 because the variation in Y_i is limited by the selection process:

$$\begin{aligned} \text{var}(\varepsilon_i | H_i = 1) &= \text{var}(\varepsilon_i | u_i > -W_i) \\ &= \sigma_\varepsilon^2 (1 - \rho^2\delta_i), \end{aligned}$$

where $\delta_i = (\lambda_i) (\lambda_i + W_i)$. We can write the regression model more compactly as

$$Y_i | (H_i = 1) = I_i + \theta \lambda_i + v_i, \quad i = 1, \dots, n_1$$

where $\theta = \rho \sigma_\varepsilon$, and $v_i = \varepsilon_i | (H_i = 1) - E(\varepsilon_i | H_i = 1)$. Note that by construction the random variable v_i has an expected value of 0.

The conditional mean value of Y_i for a selected observation is given by

$$\begin{aligned} E(Y_i | H_i = 1) &= I_i + \theta \lambda_i \\ &= I_i + \theta \frac{\phi(W_i)}{\Phi(W_i)}. \end{aligned}$$

Note that it is possible to construct a counterfactual for an observation that was not selected into the sample by observing that

$$\begin{aligned} E(\varepsilon_i | H_i = 0) &= E(\varepsilon_i | u_i \leq -W_i) \\ &= \rho \sigma_\varepsilon \frac{-\phi(W_i)}{1 - \Phi(W_i)} \\ &= -\theta \left[\frac{\phi(W_i)}{1 - \Phi(W_i)} \right] \neq 0. \end{aligned}$$

The expected value of Y_i for a nonselected observation (e.g. the expected market earnings for someone not in the labor force) is given by

$$\begin{aligned} E(Y_i | H_i = 0) &= I_i + E(\varepsilon_i | H_i = 0) \\ &= I_i - \theta \left[\frac{\phi(W_i)}{1 - \Phi(W_i)} \right]. \end{aligned}$$

Marginal effects for the selected sample depend upon whether or not the variable appears in the selection (probit) equation, in the regression equation, or in both. First, consider an explanatory variable that appears in both equations, say $Z_{ik} = X_{ik}$. The marginal effect on Y_i depends on a direct effect β_k and an indirect effect through selection γ_k :

$$\frac{\partial E(Y_i | H_i = 1)}{\partial X_{ik}} = \beta_k - \theta \gamma_k \delta_i.$$

If X_{ik} is a binary variable, the marginal effect would be

$$\begin{aligned}\Delta E(Y_i|H_i = 1)_k &= E(Y_i|H_i = 1)_{(X_{ik}=1)} - E(Y_i|H_i = 1)_{(X_{ik}=0)} \\ &= \beta_k + \theta [\lambda_{i(X_{ik}=1)} - \lambda_{i(X_{ik}=0)}].\end{aligned}$$

Next consider the marginal effect of a variable X_{ik} that appears only in the regression equation and not in the selection equation. In this case there is no indirect selection effect and the marginal effect is the same no matter whether X_{ik} is continuous or binary:

$$\frac{\partial E(Y_i|H_i = 1)}{\partial X_{ik}} = \beta_k.$$

Finally, consider the marginal effect of a variable Z_{ij} that appears only in the selection equation so that its effect on Y_i appears only through its effect on the probability of observing Y_i :

$$\frac{\partial E(Y_i|H_i = 1)}{\partial Z_{ij}} = -\theta\gamma_j\delta_i.$$

If Z_{ij} is a binary variable, the marginal effect would be

$$\begin{aligned}\Delta E(Y_i|H_i = 1)_j &= E(Y_i|H_i = 1)_{(Z_{ij}=1)} - E(Y_i|H_i = 1)_{(Z_{ij}=0)} \\ &= \theta [\lambda_{i(Z_{ij}=1)} - \lambda_{i(Z_{ij}=0)}].\end{aligned}$$

Estimation of the sample selection model by *OLS* would be biased and inconsistent unless we could observe and control for λ_i or unless $\rho = 0$ in which case there is no correlation between the selection process (u_i) and the value of Y_i (ε_i). The Heckit estimation method is a two-step procedure for obtaining consistent estimates of the regression coefficients. First, the probit model is estimated to obtain estimates of the IMR for the selected sample:

$$\hat{\lambda}_i = \frac{\phi(\hat{W}_i)}{\Phi(\hat{W}_i)},$$

where $\hat{W}_i = \hat{\gamma}_0 + \sum_{j=1}^J \hat{\gamma}_j Z_{ij}$. The probit model is estimated for the entire sample $n = n_0 + n_1$, where n_0 is the number of observation for which $H_i = 0$. The second

step is to substitute $\hat{\lambda}_i$ for λ_i in the regression model to obtain

$$Y_i | (H_i = 1) = I_i + \theta \hat{\lambda}_i + v_i^*, \quad i = 1, \dots, n_1$$

which is then estimated by *OLS*. The model could also be estimated by *MLE* but the Heckit two-step would still be estimated to furnish starting values for the parameters. The log likelihood function for the entire sample is given by

$$\begin{aligned} \ln(L) &= \sum_{H_i=0} \ln [1 - \Phi(W_i)] \\ &+ \sum_{H_i=1} \left\{ -\frac{1}{2} \left[\ln(2\pi) + \ln(\sigma_\varepsilon^2) + \frac{(Y_i - I_i)^2}{\sigma_\varepsilon^2} \right] \right. \\ &\left. + \ln \Phi \left[\left[W_i + \rho \frac{(Y_i - I_i)}{\sigma_\varepsilon} \right] \left[\frac{1}{(1 - \rho^2)^{\frac{1}{2}}} \right] \right] \right\}. \end{aligned}$$

In the Heckit model the conditional mean value of Y and marginal effects are most simply estimated as the sample average of the individual estimated effects. To begin we have the estimated conditional mean value of Y based on the mean values of the variables for individuals who were selected into the sample:

$$\overline{E(Y|H = 1)} = \bar{I} + \hat{\theta} \bar{\lambda},$$

where $\bar{I} = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k \bar{X}_k$, $\bar{X}_k = \frac{\sum_{H_i=1} X_{ik}}{n_1}$, and $\bar{\lambda} = \frac{\sum_{H_i=1} \hat{\lambda}_i}{n_1}$. The corresponding marginal effects are estimated as shown below.

First, consider an explanatory variable that appears in both equations, say $Z_k = X_k$. The marginal effect in this case is calculated according to

$$\frac{\partial \overline{E(Y|H = 1)}}{\partial X_k} = \hat{\beta}_k - \hat{\theta} \hat{\gamma}_k \bar{\delta},$$

where $\bar{\delta} = \frac{\sum_{H_i=1} \hat{\delta}_i}{n_1}$, and $\hat{\delta}_i = \left(\hat{\lambda}_i \right) \left(\hat{\lambda}_i + \hat{W}_i \right)$. If X_k is a binary variable, the estimated marginal effect would be

$$\overline{\Delta E(Y|H = 1)}_k = \overline{E(Y|H = 1)}_{(X_k=1)} - \overline{E(Y|H = 1)}_{(X_k=0)},$$

where $\overline{E(Y|H=1)}_{(X_k=1)} - \overline{E(Y|H=1)}_{(X_k=0)} = \hat{\beta}_k + \hat{\theta} \frac{\sum_{H_i=1} [\hat{\lambda}_{i(X_{ik}=1)} - \hat{\lambda}_{i(X_{ik}=0)}]}{n_1}$.

Next consider the estimated marginal effect of a variable X_k that appears only in the regression equation and not in the selection equation. In this case the estimated marginal effect is the same no matter whether X_k is continuous or binary:

$$\frac{\partial E(\widehat{Y|H=1})}{\partial X_k} = \hat{\beta}_k.$$

Finally, consider the estimated marginal effect of a variable Z_j that appears only in the selection equation:

$$\frac{\partial E(Y|H=1)}{\partial Z_j} = -\hat{\theta} \hat{\gamma}_j \bar{\delta}.$$

If Z_j is a binary variable, the estimated marginal effect would be

$$\overline{\Delta E(Y|H=1)}_j = \overline{E(Y|H=1)}_{(Z_j=1)} - \overline{E(Y|H=1)}_{(Z_j=0)},$$

where

$$\overline{E(Y|H=1)}_{(Z_j=1)} - \overline{E(Y|H=1)}_{(Z_j=0)} = \hat{\theta} \frac{\sum_{H_i=1} [\hat{\lambda}_{i(Z_{ij}=1)} - \hat{\lambda}_{i(Z_{ij}=0)}]}{n_1}.$$