

1 Qualitative Dependent Variable Models

In this section, we consider models where the dependent variable is discrete in nature. This includes qualitative choices such as mode of transportation (e.g., car, bus, train, walk, etc.); get married or stay single; teacher evaluations; number of times attending church in a year, etc.

1.1 Linear Probability Model

Consider the linear probability (LP) model

$$y_i = \beta' x_i + \mu_i$$

where $E(\mu_i) = 0$. The conditional expectation

$$E(y_i|x_i) = \beta' x_i$$

is interpreted as the probability of an event occurring given x_i . There are a couple of drawbacks to the LP model that limits its use:

1. Heteroscedasticity. Given that $y_i = \{0, 1\}$, the error term can take on two values with probability

| | |
|------------------|------------------|
| μ_i | $f(\mu_i)$ |
| $1 - \beta' x_i$ | $\beta' x_i$ |
| $-\beta' x_i$ | $1 - \beta' x_i$ |

so that the variance is

$$\begin{aligned} \text{var}(\mu_i) &= \beta' x_i (1 - \beta' x_i)^2 + (1 - \beta' x_i) (-\beta' x_i)^2 \\ &= \beta' x_i (1 - \beta' x_i) \\ &= E(y_i) [1 - E(y_i)]. \end{aligned}$$

2. Predictions outside [0,1]. The predicted probabilities from the LP model, $\hat{y}_i = \beta' x_i$, can be less than zero and greater than one.

1.2 Binomial Probit and Logit Models

The drawbacks of the LP model are solved by letting the probability of an event (i.e., $y = 1$) be given by a well-defined cumulative density function

$$Prob(y_i = 1|x) = \int_{-\infty}^{x'\beta} f(t)dt = F(x'\beta). \quad (1)$$

In this manner, the predicted probabilities will always be bounded between zero and one. If $F(x'\beta)$ is the cdf for a standard normal random variable, we get the probit model. If

$$F(x'\beta) = \frac{e^{x'\beta}}{1 + e^{x'\beta}},$$

then we get the logit model. Estimates from the logit and probit models often give similar results. The logit model is less computationally intense because $F(x'\beta)$ has a closed form, however, the logistic pdf $f(\cdot)$ has fatter tails than the standard normal pdf. Because $y_i = \{0, 1\}$ is discrete, while (1) implies continuity, we replace y_i with the latent variable y_i^* . This produces

$$y_i^* = \beta'x_i + \mu_i.$$

y_i^* can be interpreted as an unobservable index function that measures individual i 's propensity to choose $y = 1$. For example, y_i^* could be the net benefits (benefits less costs) of selecting option A. Alternatively, y_i^* could be interpreted as the difference in utility derived from choosing option A less the utility of choosing option B. Therefore, we assume

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

The choice of zero as a threshold is innocuous if the vector x_i includes a constant term.

1.2.1 Estimation

The parameters of the model are estimated via maximum likelihood. The relevant probability can be written as

$$Prob(y_i = 1|x) = Prob(y_i^* > 0|x) = Prob(\beta'x_i + \mu_i > 0|x) = Prob(\mu_i > -\beta'x_i|x).$$

Assuming a symmetric, mean-zero pdf for μ_i , we have

$$Prob(\mu_i > -\beta'x_i|x) = Prob(\mu_i < \beta'x_i|x).$$

It will be convenient to standardize μ_i , which gives

$$Prob\left(\frac{\mu_i}{\sigma} < \left(\frac{\beta}{\sigma}\right)'x_i|x\right) = \Phi\left(\left(\frac{\beta}{\sigma}\right)'x_i\right),$$

where $\Phi(\cdot)$ and σ are the cdf and standard deviation for μ_i , respectively. Therefore, the parameters are only identifiable up to a scalar σ , which is commonly set to unity (i.e., $\sigma = 1$). The likelihood function is given by

$$L = \prod_{i=1}^n [\Phi_i^{y_i} \{1 - \Phi_i\}^{1-y_i}]$$

and the log-likelihood function is given by

$$\ln L(\beta) = \sum_{i=1}^n \{y_i \ln(\Phi_i) + (1 - y_i) \ln(1 - \Phi_i)\}. \quad (2)$$

Maximization of (2) will require nonlinear optimization methods, such as Newton's algorithm.

1.2.2 Marginal Effects

The estimated coefficients, $\hat{\beta}_{ML}$, are problematic in two senses:

1. The true β s are not identified. Recall, that all we can really estimate is β/σ .
2. Aside from problem #1, we know that

$$\hat{\beta}_k = \frac{\partial y_i^*}{\partial x_{i,k}}.$$

Because y_i^* is an unobservable index function, it is difficult to interpret this derivative.

A simple solution is to calculate

$$\hat{\delta}_{i,k} = \frac{\partial Prob(y_i = 1)}{\partial x_{i,k}} = \phi\left(\left(\frac{\beta}{\sigma}\right)'x_i\right) \frac{\beta_k}{\sigma} \quad (3)$$

where $\phi(\cdot)$ is the pdf for μ_i . The advantage of the estimated marginal effect, $\hat{\delta}_{i,k}$, is that it only depends on β/σ (so that it is identifiable) and it is easy to interpret. Note that $\hat{\delta}_{i,k}$ depends on the entire vectors for x_i and β . The standard errors for $\hat{\delta}_{i,k}$ can be calculated using the delta method, which is based on a first-order Taylor approximation. We have

$$asy.var.(\hat{\delta}) = \left(\frac{\partial \hat{\delta}}{\partial \hat{\beta}'}\right) V \left(\frac{\partial \hat{\delta}}{\partial \hat{\beta}'}\right)'$$

where V is the variance-covariance matrix for $\hat{\beta}_{ML}$.

1.2.3 Goodness of Fit

Unfortunately, the standard R^2 measure of goodness of fit does not have the same interpretation (i.e., percentage of variation in Y explained by the variation in X) in binary choice models. Many alternatives have been suggested, of which a few are:

- McFadden's pseudo R^2 . This measure,

$$\tilde{R}^2 = 1 - \frac{\ln L_U}{\ln L_R},$$

is bounded between zero and one but is difficult to interpret between the limits. It is not uncommon to see low \tilde{R}^2 values (e.g., less than 0.25) for models that seemingly explain the data well.

- Likelihood ratio statistic. The standard likelihood ratio statistic is

$$LR = -2(\ln L_R - \ln L_U)$$

and is asymptotically distributed chi-square.

- Table of hits and misses. In the binary case, a 2 x 2 table can be created to summarize the number of correct and incorrect predictions. Typically, predicted probabilities greater than 0.5 (i.e., $\Phi(\beta' x_i) > 0.5$) are associated with $\hat{y}_i = 1$. The main diagonal gives the number of correct predictions and the off-diagonal gives the number of incorrect predictions.

1.2.4 Fixed and Random Effects Models for Panel Data

Sometimes we may have a panel of cross sectional - time series data intended to explain a single binary choice. Consider the following extension of the binary models above

$$\begin{aligned} y_{it}^* &= x'_{it}\beta + \mu_i + \nu_{it} \\ y_{it} &= 1 \text{ if } y_{it}^* > 0. \end{aligned}$$

As before, whether we treat μ_i as a fixed or random effect depends upon the correlation (or lack thereof) between x_{it} and μ_i . Both fixed and random effects versions of this binary choice model are available. However, there are additional complications above and beyond the standard quantitative dependent-variable cases. In particular, in the RE case, the likelihood function will involve integration over the μ_i and, in the FE case, it is not possible to remove the fixed-effects term μ_i by subtracting group means. See section 21.5.1 in Greene for more details.

1.2.5 Bivariate Probit Model

The bivariate binary choice model takes the (SUR-like) form

$$\begin{aligned} y_1^* &= x_1' \beta_1 + \epsilon_1 \\ y_2^* &= x_2' \beta_2 + \epsilon_2 \end{aligned}$$

where $E(\epsilon_1) = E(\epsilon_2) = 0$ and

$$\text{var}(\epsilon) = \text{var}[(\epsilon_1 \ \epsilon_2)'] = E(\epsilon\epsilon') = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

As before, y_1^* and y_2^* are unobserved index functions such that

$$\begin{aligned} y_1 &= 1 \text{ if } y_1^* > 0 \text{ and } y_1 = 0 \text{ otherwise;} \\ y_2 &= 1 \text{ if } y_2^* > 0 \text{ and } y_2 = 0 \text{ otherwise.} \end{aligned}$$

There are four possible outcomes in the binary case. For example, the probability that $y_{i1} = y_{i2} = 0$ is

$$P_{00} = \text{Prob}(y_{1i} = 0, y_{2i} = 0 | x_1, x_2) = \int_{-\infty}^{-x_1' \beta_1} \int_{-\infty}^{-x_2' \beta_2} \phi(\epsilon_1, \epsilon_2; \rho) d\epsilon_1 d\epsilon_2$$

where $\phi(\epsilon_1, \epsilon_2; \rho)$ represents the bivariate pdf. If $\phi(\epsilon_1, \epsilon_2; \rho)$ is specified as the bivariate normal pdf, then we have the bivariate probit model. The log likelihood function is

$$\ln L(\beta_1, \beta_2, \rho) = \sum_{y_1=0, y_2=0} \ln P_{i,00} + \sum_{y_1=1, y_2=0} \ln P_{i,10} + \sum_{y_1=0, y_2=1} \ln P_{i,01} + \sum_{y_1=1, y_2=1} \ln P_{i,11}$$

which is maximized through nonlinear optimization methods by choosing $\theta = (\beta_1, \beta_2, \rho)$. A potentially useful test is the Lagrange multiplier test to see whether $\rho = 0$ so that the probit models can be estimated separately (see Greene section 21.6.2). Marginal effects can be calculated (although they are a bit more complicated than the univariate probit case) and the delta method can be used to calculate standard errors.

1.3 Multinomial (Unordered) Logit

Consider explaining J different unordered choices (e.g., religion choice – Protestant, Catholic, Islam, Hindu, etc.). The multinomial logit model is derived by letting

$$\text{Prob}(y_i = j) = \frac{\exp(\beta_j' x_i)}{\sum_{k=1}^J \exp(\beta_k' x_i)}$$

for $j = 1, 2, \dots, J$.¹ Typically the normalization $\beta'_1 = 0$ is made, which gives

$$P_{ij} = Prob(y_i = j) = \frac{\exp(\beta'_j x_i)}{1 + \sum_{k=2}^J \exp(\beta'_k x_i)}$$

and clearly satisfies the condition that $P_{i1} + P_{i2} + \dots + P_{iJ} = 1$ for all $i = 1, \dots, n$. This model gets the name multinomial logit because we assume that the binary probability

$$\frac{P_{ij}}{P_{i1} + P_{ij}} = \frac{\exp(\beta'_j x_i)}{1 + \exp(\beta'_j x_i)}$$

is given by the logistic cdf for $j = 2, \dots, J$. The likelihood function is

$$L(\beta_2, \beta_3, \dots, \beta_J) = \left[\prod_{y_i=1} P_{i1} \right] \left[\prod_{y_i=2} P_{i2} \right] \dots \left[\prod_{y_i=J} P_{iJ} \right]. \quad (4)$$

Maximization of (4) will produce $J - 1$ coefficient vectors. The marginal effects are

$$\delta_{ij} = \frac{\partial P_{ij}}{\partial x_i} = P_{ij}[\beta_j - \bar{\beta}]$$

where $\bar{\beta}$ is the average coefficient vector. Therefore, the marginal effects for the j^{th} choice depends upon i and the parameters for all the choices $(\beta_2, \beta_3, \dots, \beta_J)$. Standard errors can be calculated through the delta method (Greene, p. 722).

As a final note, the log-odds ratio is

$$\ln(P_{ij}) - \ln(P_{i1}) = x'_i \beta_j$$

and only depends upon β_j . This is called the independence of irrelevant alternatives (IIA) and it is a feature of the multinomial logit model. To test for IIA, Hausman and McFadden provide the following test statistic

$$HM = (\hat{\beta}_R - \hat{\beta}_U)' [\hat{V}_R - \hat{V}_U]^{-1} (\hat{\beta}_R - \hat{\beta}_U) \stackrel{asy}{\sim} \chi^2(K).$$

where the R subscript denotes the model with the other choices omitted and U denotes the full model. Should the HM statistic indicate a rejection of the null hypothesis of IIA, then the disturbances may not be independent and homoscedastic. In this case, one alternative to the multinomial logit model is a multivariate model (such as the bivariate case described above), which allows for correlations across alternatives. Another alternative is the nested logit model, where we break the choices into subgroups, where the IIA may hold within a group but not across subgroups. An example is the choice of community and type of housing, which

¹When x_{ij} consists of choice-specific, as opposed to individual-specific characteristics, the appropriate model is the conditional logit model. The conditional logit differs from the multinomial logit model in that the coefficients do not vary across choices.

can be nested by first considering the choice of community and then the choice of housing type, conditional on the chosen community. See section 21.7.4 in Greene for more details.

1.4 Ordered Probit

Consider the following index function model

$$y_i^* = \beta' x_i + \mu_i$$

used to explain the ordered choices $y_i = \{1, 2, \dots, m\}$. One example is choice of educational level, such as high school degree ($y_i = 1$), undergraduate degree ($y_i = 2$) or graduate degree ($y_i = 3$). We assume that

$$\begin{aligned} \text{if } y_i^* < \alpha_1 & \text{ then } y_i = 1 \\ \text{if } \alpha_1 < y_i^* < \alpha_2 & \text{ then } y_i = 2 \\ \text{if } \alpha_2 < y_i^* < \alpha_3 & \text{ then } y_i = 3 \\ & \vdots \\ \text{if } y_i^* > \alpha_{m-1} & \text{ then } y_i = m \end{aligned}$$

where the α_j are the threshold values for $j = 1, 2, \dots, m - 1$. Probabilities are given by

$$\begin{aligned} P_1 &= \Phi(\alpha_1 - \beta' x_i) \\ P_2 &= \Phi(\alpha_2 - \beta' x_i) - \Phi(\alpha_1 - \beta' x_i) \\ P_3 &= \Phi(\alpha_3 - \beta' x_i) - \Phi(\alpha_2 - \beta' x_i) \\ &\vdots \\ P_m &= 1 - \sum_{j=1}^{m-1} P_j. \end{aligned}$$

The likelihood function is given by

$$L(\beta) = \left[\prod_{y_i=1} P_{1,i} \right] \left[\prod_{y_i=2} P_{2,i} \right] \cdots \left[\prod_{y_i=m} P_{m,i} \right]. \quad (5)$$

Maximum likelihood estimates are found by taking natural logs of (5) and then maximizing by choosing β . Note that although there are m different choices, there is only a single coefficient vector β . The ordered probit model results if Φ_i is specified as the standard normal cdf. Again, this will result in a nonlinear optimization problem.

1.5 Count Data

Count data refers to a dependent variable y_i that takes on values from the set $\{0, 1, 2, \dots\}$. For example, the number of automobile accidents per year is an example of count data. A common way to model this type of environment is through a Poisson regression model, such that each y_i is drawn from the Poisson distribution. Recall that the Poisson pdf is

$$Prob(Y_i = y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

with the property that $E(Y_i) = var(Y_i) = \lambda_i$ for $y_i = 0, 1, 2, \dots$. In order to make this into a regression model, we specify that

$$\ln \lambda_i = x_i' \beta$$

such that the expected number of events per period is

$$E[Y_i|x_i] = \lambda_i = \exp(x_i' \beta).$$

Rather than report the estimates of β directly, it makes more sense to report the marginal effects

$$\frac{\partial E[Y_i|x_i]}{\partial x_i} = \lambda_i \beta.$$

This model is typically estimated using maximum likelihood and the (log) likelihood function is

$$\ln L(\beta) = \sum_{i=1}^n [-\lambda_i + y_i x_i' \beta - \ln y_i!].$$

A common test in the case of the Poisson regression model is a test for overdispersion (i.e., test whether the mean equals the variance). If this test indicates that the mean and variance are not equal, then the Poisson distribution may be inappropriate and an alternative such as the negative binomial model might be specified.

2 Limited Dependent Variable Models

2.1 Truncation

Truncation models refer to cases where only a subset of the population is observed. For example, a mail survey sent to residents of a community is likely to exclude homeless people. In this instance, the distribution on income is likely to be truncated from below.

2.1.1 Truncated Distributions and Moments

A truncated distribution is a conditional distribution. For example, the probability density function (pdf) for x when it is truncated from below at point $x = a$ is

$$f(x|x > a) = \frac{f(x)}{Prob(x > a)} = \frac{f(x)}{1 - F(a)}$$

where $F(x)$ is the cumulative distribution function (cdf) for random variable x . In the case of the standard normal pdf $\phi(z)$, the truncated distribution is

$$f(x|x > a) = \frac{\sigma^{-1}\phi(z)}{1 - \Phi(\alpha)} = \frac{\sigma^{-1}[(2\pi)^{-0.5} \exp(-0.5z^2)]}{\int_{\alpha}^{\infty} \phi(z) dz}$$

where $\alpha = \sigma^{-1}(a - \mu)$ and $z = \sigma^{-1}(x - \mu)$. The mean and variance of $f(x|x > a)$ are

$$\mu_{x|x>a} = \mu + \sigma\lambda(\alpha) \tag{6}$$

$$\sigma_{x|x>a}^2 = \sigma^2[1 - \delta(\alpha)] \tag{7}$$

where $\lambda(\alpha)$ is the inverse Mills ratio

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)] > 0 \text{ when truncated from below;}$$

$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha) < 0 \text{ when truncated from above}$$

and

$$0 < \delta(\alpha) = \lambda(\lambda - \alpha) < 1.$$

The size of the truncated mean and variance formula (relative to the standard case) make sense. Consider truncation from below so that $\lambda(\alpha) > 0$. As expected, because $\sigma\lambda(\alpha) > 0$, truncation increases the mean. Also as expected, because $0 < \delta < 1$, truncation decreases the variance.

2.1.2 Truncated Regression Model

Begin with our standard linear regression model

$$y_i = x_i'\beta + \epsilon_i$$

where $y_i|x_i \sim N[x_i'\beta, \sigma^2]$ when y_i is not truncated. Now assume that we are only interested in the portion of the distribution where $y_i > a$. Next, we decompose y_i into its deterministic and stochastic parts

$$y_i = E[y_i|y_i > a] + \nu_i$$

which after using equation (6) gives

$$y_i = x_i'\beta + \sigma\lambda_i + \nu_i. \tag{8}$$

Note also that $E(\nu_i) = 0$ and $Var(\nu_i) = \sigma^2(1 - \delta_i)$ so that equation (8) suffers from heteroscedasticity. Consequently, OLS of y on x will be inefficient. Furthermore, because of the term λ_i , it will also suffer from omitted-variable bias. Maximum likelihood estimation, incorporating the term λ_i , is the standard approach and will provide consistent estimates.

2.2 Censoring and the Tobit Model

Censoring is related to truncation, but it is not quite the same. Under truncation, we do not observe any $y_i < a$, whereas under censoring, all $y_i < a$ are set at $y_i = a$. One example is the quantity of tickets demanded for UW football games. For games that are not sold out, we can use the actual attendance as a measure of demand. For games that are sold out, however, demand is censored at the capacity of War Memorial stadium. To get an accurate prediction of ticket demand, the censored nature of the data needs to be taken into account.

2.2.1 Censoring

Under censoring, we have

$$\begin{aligned} y &= a \text{ if } y^* \leq a \\ y &= y^* \text{ if } y^* > a. \end{aligned}$$

Therefore, y is a random variable with a mixture of discrete and continuous probability distributions. When $y^* > a$, the distribution is, for example, $y^* \sim N(\mu, \sigma^2)$. When $y^* \leq a$, the distribution is $Prob(y = a) = Prob(y^* \leq a) = \Phi(\alpha)$, where $\alpha = (a - \mu)/\sigma$. The mean and variance of the censored distribution are

$$\begin{aligned} \mu_y &= \Phi(\alpha)a + [1 - \Phi(\alpha)](\mu + \sigma\lambda) \\ \sigma_y^2 &= \sigma^2[1 - \Phi(\alpha)][(1 - \delta) + (\alpha - \lambda)^2\Phi(\alpha)] \end{aligned}$$

where λ and δ are as defined above. If $\Phi(\alpha) = 0$, so there is no censoring, then we simply get the truncated mean and variance.

2.2.2 Censored Regression (Tobit) Model

If we make the additional assumption that $a = 0$ and

$$y_i^* = x_i' \beta + \epsilon_i,$$

then we have the censored regression (or Tobit) model. The expected values are

$$\begin{aligned} E[y_i^* | x_i] &= x_i' \beta \\ E[y_i | x_i] &= \Phi\left(\frac{x_i' \beta}{\sigma}\right) [x_i' \beta + \sigma \lambda_i] \end{aligned}$$

where the inverse Mills ratio is

$$\lambda_i = \frac{\phi(\sigma^{-1} x_i' \beta)}{\Phi(\sigma^{-1} x_i' \beta)}.$$

The model can be estimated using two-step procedures (details under sample selection) or maximum likelihood. Assuming normally distributed errors, the log likelihood function ($z_i = \sigma^{-1}(y_i - x_i' \beta)$) can be written as

$$\ln L(\beta, \sigma) = \sum_{y_i > 0} \ln[\sigma^{-1} \phi(z_i)] + \sum_{y_i = 0} \ln[1 - \Phi(\sigma^{-1} x_i' \beta)]$$

and maximized with respect to β and σ . All the desirable maximum likelihood results apply. Marginal effects are more revealing than the parameter estimates themselves. Marginal effects are

$$\begin{aligned} \frac{\partial E[y_i^* | x_i]}{\partial x_i} &= \beta \\ \frac{\partial E[y_i | x_i]}{\partial x_i} &= \beta \times \text{Prob}(y_i^* > a). \end{aligned}$$

2.3 Sample Selection

In this section, we consider models where the sample is non-random. Examples include:

- Survey design. For example, if we are interested in the effect of 401K participation on family wealth but we only have access to households with wealth greater than \$100,000, then the sample is non-random. Another example would be surveying households about their maximum willingness to pay for a community program, but the survey only covers communities where the program is already in

place.

- Incidental Truncation. Incidental truncation refers to truncation of a sample due to some other variable. For example, estimating a wage offer equation may exhibit incidental truncation if we only observe individuals that select to participate in the labor market.

2.3.1 Bivariate Incidental Truncation

Let y and z have bivariate normal distribution with distribution $f(y, z; \rho)$. Now assume we that we only observe y when $z > a$. The truncated joint density is

$$f(y, z|z > a) = \frac{f(y, z)}{\text{prob}(z > a)}.$$

The mean of $f(y, z|z > a)$ is

$$E(y|z > a) = \mu_y + \rho\sigma_y\lambda\left(\frac{a - \mu_z}{\sigma_z}\right)$$

where $\lambda(\cdot) = \phi(\cdot)/[1 - \Phi(\cdot)]$ is the inverse Mills ratio.

2.3.2 Sample Selection Model

Assume the sample is generated from the joint system

$$z_i^* = w_i'\gamma + u_i \quad (\text{Participation Equation})$$

$$y_i = x_i'\beta + \epsilon_i \quad (\text{Main Equation})$$

where y_i is observed when $z_i^* > 0$. Also, suppose that u_i and ϵ_i have a bivariate normal distribution

$$f(u, \epsilon; \rho)$$

and zero means. For estimation purposes, we decompose the observed y_i into its deterministic and stochastic components:

$$y_i|z_i^* > 0 = E[y_i|z_i^* > 0] + v_i$$

where

$$\begin{aligned}
E[y_i | z_i^* > 0] &= E[y_i | w_i' \gamma + u_i > 0] \\
&= E[y_i | u_i > -w_i' \gamma] \\
&= x_i' \beta + E[\epsilon_i | u_i > -w_i' \gamma] \\
&= x_i' \beta + \rho \sigma_\epsilon \lambda_i \left(\frac{-w_i' \gamma}{\sigma_u} \right) \\
&= x_i' \beta + \beta_\lambda \lambda_i \left(\frac{-w_i' \gamma}{\sigma_u} \right).
\end{aligned}$$

Therefore, an OLS regression of y_i on only x_i will result in an inconsistent estimate of β . An OLS regression of y_i on x_i and λ_i would produce consistent estimates of β (albeit inefficient because v_i is heteroscedastic).

2.3.3 Marginal Effects

The relevant marginal effect is

$$\frac{\partial E[y_i | z_i^* > 0]}{\partial x_{ik}} = \beta_k - \frac{\rho \sigma_\epsilon \gamma_k \lambda_i}{\sigma_u} \left[\lambda_i + \frac{w_i' \gamma}{\sigma_u} \right].$$

This marginal effect has two parts – a direct effect of x_k on y (i.e., β_k) and an indirect effect of x_k on y through the participation equation.

2.3.4 Estimation

The parameters of the selection model can be estimated through direct maximum likelihood methods or through a two-step procedure. The two step procedure, credited to Heckman (1979) and often called the "Heckit" method is as follows:

- Step #1. Using probit, estimate the participation equation and obtain $\hat{\gamma}/\sigma_u$. Use this to calculate the inverse Mills ratio for each observation, $\hat{\lambda}_i$.
- Step #2. Using OLS, estimate the main equation with a regression of y on x and $\hat{\lambda}$.

This produces consistent estimates of the parameters. It is also possible to recover estimates for ρ and σ_ϵ separately.

2.3.5 An Application: The effects of ACT scores on college performance

Consider the effects of ACT scores on academic performance in college. Let the primary equation be

$$SCORE_i = \beta_0 + \beta_1 ACT_i + \epsilon_i$$

where $SCORE$ is a performance measure for college on a scale from 0 to 100, ACT measures scores on the college-entrance exam and ϵ is a normally distributed error term. The goal is to estimate the direct effect of ACT scores on performance in college. OLS estimates $\hat{\beta}_1$ are likely to incorrectly estimate the direct effect because a regression of SCORE on ACT for college students does not account for the effect that ACT has on college admission/enrollment.

Consider the college participation equation

$$COLLEGE_i^* = \gamma_0 + \gamma_1 ACT_i + u_i$$

where $COLLEGE^*$ is the latent indicator variable for college and u is a normally distributed random variable that is positively correlated (ρ) with ϵ . The positive correlation picks up the notion that those who are most likely to go to college are also most likely to perform well in college. As before, we impose that $COLLEGE_i = 1$ when $COLLEGE_i^* > 0$ and $COLLEGE_i = 0$ when $COLLEGE_i^* \leq 0$.

Using the incidental truncation results above, we know that

$$\begin{aligned} E[SCORE_i | COLLEGE_i = 1] &= \beta_0 + \beta_1 ACT_i + E(\epsilon_i | COLLEGE_i = 1) \\ &= \beta_0 + \beta_1 ACT_i + \rho \sigma_\epsilon \lambda_i \end{aligned}$$

where $\lambda_i(\gamma_0 + \gamma_1 ACT_i) = \frac{\phi(\gamma_0 + \gamma_1 ACT_i)}{\Phi(\gamma_0 + \gamma_1 ACT_i)}$ is the (left-truncated) inverse Mills ratio. Naive OLS estimates suffer from omitted-variable bias because they omit λ_i . The two-stage Heckit procedure (whereby we estimate λ_i in the first stage and then regress SCORE on a constant, ACT and λ_i), however, will produce consistent estimates of β_1 .