Binary Choice Models

Table of contents

1 Overview

In this handout we will review the binary case of discrete choice models (a type of limited dependent variable model). We will review:

- linear probability model (LPM);
- probit model;
- and logit model.

Studying the simple case binary choice models, will provide you with tools needed to explore richer models like multinomial logit/probit, ordered logit/probit, and conditional logit.

Further reading can be found in:

- Section 14-14.4 of Cameron and Trivedi (2005)
- Section 7-7.1.6 of Verbeek (2017)

2 Binary Outcomes

Consider a Bernoulli random variable $Y \in \{0, 1\}$. Given the vector of random variables $X \in \mathbb{R}^k$, we can define

$$\rho(x) = Pr(Y = 1|X = x)$$

and,

$$1 - \rho(x) = 1 - Pr(Y = 1|X = x) = Pr(Y = 0|X = x)$$

Then, the conditional mean is given by,

$$E[Y|X] = 0 \times Pr(Y = 0|X) + 1 \times Pr(Y = 1|X) = \rho(X)$$

Likewise, the conditional variance is,

$$\begin{aligned} Var(Y|X) =& E[Y^2|X] - E[Y|X]^2 \\ =& E[Y|X] - E[Y|X]^2 \\ =& \rho(X) - \rho(X)^2 \\ =& \rho(X)(1 - \rho(X)) \end{aligned}$$

where the second line follows from the fact that $Y = Y^2$.

3 Linear Probability Model

The LPM is simply a standard linear model with a binary outcome:

$$Y_i = X_i'\beta + \varepsilon_i$$

where $Y_i \in \{0, 1\}$. We can rationalize this model by assuming that the conditional expectation function of Y is linear (in parameters),

$$E[Y_i|X_i] = X_i'\beta = Pr(Y_i = 1|X_i)$$

However, since the conditional expectation is equivalent to the conditional probability (that $Y = 1$), it must be that:

$$0 \leq X_i'\beta \leq 1$$

Outside of this interval, the model is not defined.

Recall, assuming knowledge of the CEF is equivalent to assuming conditional mean independence of the error term, since we can always write $Y_i = E[Y_i|X_i] + \varepsilon_i$ where $E[\varepsilon_i|X_i] = 0$. In this setting, the error term can take on 2 values for any given value of X_i:

$$\varepsilon_i = \begin{cases} 1 - X_i'\beta & \text{for} \quad Y_i = 1 \\ -X_i'\beta & \text{for} \quad Y_i = 0 \end{cases}$$

This implies that the error cannot be normally distributed (conditional on X_i). This is limiting in terms of the finite sample distribution of estimators for β. However, the asymptotic distribution of estimators will remain normal under CLT.

What of the conditional variance? For continuous outcomes, we could assume homoskedasticity: $Var(\varepsilon_i|X_i) = \sigma^2$. Now, with a binary outcome,

$$Var(\varepsilon_i|X_i) = Var(Y_i|X_i) = \rho(X_i)(1 - \rho(X_i))$$

Thus, the error term is by definition heteroskedastic.

3.1 Estimation

The LPM can be estimated by OLS. So long as,

$$0 \leq X_i'\beta \leq 1$$

OLS is unbiased and consistent. If not, OLS is neither unbiased nor consistent.

Regardless, OLS is inefficient, given the heteroskedasticity of the error term. You should therefore estimate heteroskedastic robust SEs. An alternative is to estimate a Weighted Least Squares model.

3.2 Predicted values

An issue with the LPM model is predictions outside of the range $[0, 1]$. For example,

This usually occurs in instances where X has a large support, with the majority of observations close to the mean (of X). For example, if X is normally distributed.

4 Latent Variable Model

A latent variable model is one way of justifying the assumptions underlying the probit and logit models; however, it is not strictly required. This approach generalizes the model in a way that restricts the outcome (to a binary outcome), either conditionally or unconditionally.

We observe $Y_i \in \{0, 1\}$. We can express this outcome as a function of $\{X, \varepsilon\}$ in the following way:

$$Y_i = \mathbf{1}\{X_i'\beta + \varepsilon_i > 0\}$$

Here, $\mathbf{1}\{\cdot\}$ is an indicator function: equal to 1 when the statement is true and 0 otherwise. We can then write this as,

$$Y_i = \mathbf{1}\{Y_i^* > 0\} \qquad \text{where} \quad Y_i^* = X_i'\beta + \varepsilon_i$$

We refer to Y^* as a *latent* variable. It is unobserved; hence, the name 'latent' which can be interpretted as hidden or concealed.

There are a couple of important properties of this set-up:

1. The observed outcome is unique up to a scalar multiplication of the latent variable.

$$Y_i = \mathbf{1}\{X_i'\beta + \varepsilon_i > 0\} = \mathbf{1}\{aX_i'\beta + a\varepsilon_i > 0\}$$

This implies that the absolute magnitude of β cannot be identified. We will see this morning clearly in our discussion of the probit model.

2. The observed outcome is unique up to the addition and substraction of a constant:

$$Y_i = \mathbf{1}\{X_i'\beta + \varepsilon_i > 0\} = \mathbf{1}\{X_i'\beta + c + \varepsilon_i - c > 0\}$$

This implies that the threshold at which decisions are made cannot be identified and need not be zero.

Practically, these two probabilities mean that we will need to normalize the location (mean) and variance of the unobserved error (ε).

4.1 Utility maximization

Such a latent variable framework should be familiar to an Economics student. Afterall, von-Neumann and Morgenstern's theory of utility maximization states that under certain axioms (concerning the preferences of the individual) individual choices are consistent with the maximization of a continuously differentiable utility function. We observe individual choices (which are often discrete), not their utility function: a latent variable determining their choices.

Consider the choice between two goods/options $j = 0, 1$; for example, travel by car (=1) or bus (=0). Suppose the utility derived from each choice is given by,

$$U_{ij} = S_{ij} + \varepsilon_{ij}$$

where S_ij is a deterministic component, and ε_{ij} stochastic. For example, in our example of travel options the deterministic component could be the known cost of each option, while the stochastic component could be due to unexpected variation in travel times if a private vehicle is more affected by traffic.

Assuming utility maximization, the individual chooses to travel by car if $U_{i1} > U_{i0}$.

$$\begin{aligned} Pr(Y_i = 1) =& Pr(U_{i1} > U_{i0}) \\ =& Pr(S_{i1} + \varepsilon_{i1} > S_{i0} + \varepsilon_{i0}) \\ =& Pr(\varepsilon_{i0} - \varepsilon_{i1} < S_{i1} - S_{i0}) \end{aligned}$$

Given known values of the deterministic components $S_{i1} - S_{i0}$, we could component this probability given the CDF of $\varepsilon_{i0} - \varepsilon_{i1}$. This, of course requires a an assumption.

Suppose that the deterministic component depended on option-specific variables ($W_{ij}'\lambda$; e.g. cost of transport option) and individual-specific variables with option-specific marginal utility ($Z_i'\gamma_j$; e.g. age).

$$S_{ij} = W_{ij}'\lambda + Z_i'\gamma_j$$

Then, the probability of observing choice $j = 1$ (i.e. personal vehicle) is given by,

$$Pr(Y_i = 1|W_i, Z_i) = Pr(\varepsilon_{i0} - \varepsilon_{i1} < (W_{i1} - W_{i0})'\lambda + Z_i'(\gamma_1 - \gamma_0))$$

Suppose, $var(\varepsilon_i|W_i, Z_i) = \sigma^2$. Then we can, we can define,

$$Pr(Y_i = 1|W_i, Z_i) = Pr(\varepsilon_i < \Delta W_i'\eta + Z_i'\gamma)$$

where,

$$\varepsilon_i = \frac{\varepsilon_{i0} - \varepsilon_{i1}}{\sigma};\ \eta = \frac{\delta}{\sigma};\ \text{and } \gamma = \frac{\gamma_1 - \gamma_0}{\sigma}$$

As a result we have a $Pr(Y_i = 1|X_i) = Pr(\varepsilon_i < X_i'\beta)$ where $X_i'\beta = \Delta W_i'\eta + Z_i'\gamma$. This normalization of the parameters is standard in these models. Moreover, we cannot identify the option-specific parameters (γ_j) on the invariant variables.

Given a choice of the distribution of ε_i,

$$Pr(Y_i = 1|X_i) = F_\varepsilon(X_i'\beta)$$

4.2 Probit

The probit model assumes that the error term from the expression,

$$Y_i = \mathbf{1}\{X_i'\beta + \varepsilon_i > 0\}$$

is independently and identicially distributed (conditional on X),

$$\varepsilon_i|X_i \sim N(0, 1)$$

Note, if we assume that the conditional variance is σ^2, then we have to normalize the model since only β/σ is identified.

The PDF of the standard normal distribution is given by,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} exp(-z^2/2)$$

and the CDF by,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} exp(-u^2/2)du$$

This latter expression has no closed form solution.

The conditional probability of $Y_i = 1$ (expectation of Y_i), is given by,

$$\begin{aligned} Pr(Y_i = 1|X_i) =& Pr(X_i'\beta + \varepsilon_i > 0|X_i) \\ =& Pr(\varepsilon_i > -X_i'\beta|X_i) \\ =& Pr(\varepsilon_i < X_i'\beta|X_i) \\ =& \Phi(X_i'\beta) \end{aligned}$$

where line 3 follows from the symmetry of $N(0,1)$.

4.3 Logit

In this case, we assume that the conditional distribution of the error term is logistic. The CDF of the (standard) logistic distribution is given by,

$$\Lambda(z) = \frac{exp(z)}{1 + exp(z)} = \frac{1}{1 + exp(-z)}$$

and the PDF by,

$$\lambda(z) = \Lambda(z)(1 - \Lambda(z))$$

Like the standard normal distribution, the standard logistic distribution is symmetric. It's variance is $\pi^2/3$ and has slightly thicker tails compared to the standard normal.

4.4 LPM

If we assume that the conditional distribution of the error is uniform with mean 0 and variance=1,[1] then

$$Pr(Y_i = 1|X_i) = X_i'\beta$$

Thus, one way of justifying the use of a linear probability model is through assumption of a uniformly distributed error.

[1] Note, this is not standard uniform distribution: $U(0,1)$. A standard uniform distribution has mean 1/2 and variance 1/12.

5 Index Model Approach

An alternative approach is to consider the problem of the modeling the conditional probability (expectation) function. In the case of the LPM we had,

$$Pr(Y_i = 1|X_i) = X_i'\beta$$

More generally, we can place this **linear index** - $X_i'\beta$ - inside a function $G(\cdot)$.

$$Pr(Y_i = 1|X_i) = G(X_i'\beta)$$

The covariates affect the conditional probability only through the linear index, which is then mapped into a response probability by the function $G(\cdot)$.

For $G(\cdot)$ linear, we have the LPM. However, we know that this model can give predicted probabilities outside of the bounds of $[0, 1]$. An ideal candidate for $G(\cdot)$ would be any function that bounds the values of the predicted probabilities within $[0, 1]$. Given that a CDF is a function with values within this range, the functions $\Phi(\cdot), \Lambda(\cdot)$ are prime candidates. However, non-symmetric functions also exist; for example, the Gumbel (extreme-value or complementary log-log),

$$1 - exp(-exp(z))$$

6 Estimation

We will examine the estimation of these models by Maximum Likelihood (ML). Recall from Handout 3, we defined the ML estimator as the maximizer of the conditional log-likelihood function. Assuming an iid sample, this is given by,

$$\hat{\theta}^{ML} = \arg\max_{\theta} n^{-1} \log L_n(\theta) = \arg\max_{\theta} n^{-1} \sum_{i=1}^{n} \log \ell_i(\theta)$$

where $ell_i(\theta) = f(Y_i|X_i;\theta)$. In this application, the data $W_i = [Y_i, X_i']$ where $Y_i \in \{0, 1\}$ and $X_i \in \mathbb{R}^k$. The population parameters are $\theta = \beta$; since, we normalize $\sigma = 1$. Since, the outcome is discrete, the conditional likelihood function is given by two probabilities:

- $Pr(Y_i = 1|X_i) = F(X_i'\beta)$
- $Pr(Y_i = 0|X_i) = 1 - F(X_i'\beta)$

where $F(\cdot)$ is the CDF of the unobserved error. Hence, the joint (conditional) likelihood is given by,

$$L_n(\beta) = \prod_{i:Y_i=1} F(X_i'\beta) \prod_{i:Y_i=0} (1 - F(X_i'\beta))$$

which can be written as,

$$L_n(\beta) = \prod_{i=1}^{n} F(X_i'\beta)^{Y_i} (1 - F(X_i'\beta))^{1-Y_i}$$

The ML estimator is then by given by,

$$\hat{\beta}^{ML} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} Y_i \times \ln\left(F(X_i'\beta)\right) + (1 - Y_i) \times \ln\left(1 - F(X_i'\beta)\right)$$

Solving for the first-order conditions (FOCs), we get:

$$\frac{1}{n} \frac{\partial L_n(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \frac{f(X_i'\beta)}{F(X_i'\beta)} - (1 - Y_i) \frac{f(X_i'\beta)}{1 - F(X_i'\beta)} \right] X_i$$

The term in the square parenthesis is referred to as the **generalized error**. In this was, the above **score function** resembles the FOCs from a linear model. The generalized error can be solved for by taking the derivative of $L_n(\theta)$ with-respect-to the constant term in the model (assuming there is one).

$$\text{generalized error} = \begin{cases} \frac{f(X_i'\beta)}{F(X_i'\beta)} & \text{for} \quad Y_i = 1 \\ \frac{-f(X_i'\beta)}{1-F(X_i'\beta)} & \text{for} \quad Y_i = 0 \end{cases}$$

The scaling by $1/n$ is no impact on the solution to the FOCs, but is needed to ensure consistency of the variance-covariance matrix. This was the case for the linear model as well.

Assuming a concave likelihood function, there is a unique maximum and we can solve for $\hat{\beta}$ by setting the score function $= 0$.

6.1 Probit

In the case of the probit model, $F(z) = \Phi(z)$. Solving the FOCs, we get

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left[Y_i \frac{\phi(X_i'\beta)}{\Phi(X_i'\beta)} - (1 - Y_i) \frac{\phi(X_i'\beta)}{1 - \Phi(X_i'\beta)} \right] X_i \\ &= \sum_{i=1}^{n} \left[\frac{Y_i - \Phi(X_i'\beta)}{\Phi(X_i'\beta)(1 - \Phi(X_i'\beta))} \right] X_i \cdot \phi(X_i'\beta) \end{aligned}$$

6.2 Logit

In the case of the logit model, $F(z) = \Lambda(z)$ and $f(z) = \Lambda(z)(1 - \Lambda(z))$. This yields a simpler score function:

$$0 = \sum_{i=1}^{n} \Big[Y_i - \Lambda(X_i'\beta) \Big] X_i$$

However, neither the probit nor logit has an analytical solution and both must be solved numerically.

6.3 Asymptotic distribution

For both probit and logit, the estimators are asymptotically normal,

$$\sqrt{n}(\hat{\beta}_n^{ML} - \beta_0) \rightarrow_d N(0, V)$$

where β_0 is the true value of β,

$$V = (J(\beta_0))^{-1} = \Bigg(\underbrace{- E\Big[\frac{\partial^2}{\partial\beta\partial\beta'} \ln f(W_i, \beta_0) \Big]}_{J(\beta_0)} \Bigg)^{-1}$$

The variance is given by the inverse of the Jacobian matrix, evaluating at the true value of β. The Jacobian matrix is the (negative) expected value of the Hessian matrix of second derivatives. This result derives from the Delta Method which is used to explain the asymptotic distribution of the ML estimator.

Let us prove this result. In Handout 3 we established the consistency of the ML-estimator. To prove its asymptotic normality, we will need to use the Mean Value Theorem (MVT).[2]

Theorem 6.1 (Mean Value Theorem)**.** *For $f(\cdot)$ continuous on $[a, b]$, and continuously differentiable on (a, b), $\exists\, c \in (a, b)$ s.t.,*

$$f'(c) = \frac{f(a) - f(b)}{a - b}$$

Using MVT, we can prove asymptotic normality and demonstrate the Jacobian-variance result.

[2]The Mean Value Theorem looks very similar to the Taylor Series approximation: $f(a) \approx f(b) + f'(b)(a - b)$. When $f'(\cdot)$ is evaluated at the mean-value, c, the Taylor Series approximation is an equality.

Proof. Consider the F.O.C.s from the ML estimator, evaluated at $\hat{\beta}_n^{ML}$.

$$0 = n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log \underbrace{f(Y_i|X_i; \hat{\beta}_n)}_{\ell_i(\hat{\beta}_n)}$$

Apply MVT to the F.O.C.,

$$\begin{aligned} 0 =& n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log \ell_i(\beta_0) + n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_n^*)(\hat{\beta}_n - \beta_0) \\ \Rightarrow \sqrt{n}(\hat{\beta}_n - \beta_0) =& \left(n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_n^*) \right)^{-1} n^{-1/2} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log \ell_i(\beta_0) \end{aligned}$$

where β_n^* is a mean-value between the estimator, $\hat{\beta}_n$, and the true value of the parameter, β_0. Since, $\hat{\beta}_n \to_p \beta_0$ and

$$|\hat{\beta}_n - \beta_0| > |\beta_n^* - \beta_0|$$

by definition of a mean value, then $\beta_n^* \to_p \beta_0$. The mean-value will converge in probability to the true value as $n \to \infty$.

Then by WLLN and Slutzky's Theorem,

$$\left(n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_n^*) \right)^{-1} \to_p \left(E\left[\frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_0) \right] \right)^{-1}$$

Consider the second term,

$$\begin{aligned} & E\left[n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log \ell_i(\beta_0) \right] \\ =& E\left[\frac{\partial}{\partial \beta} \log f(Y_i|X_i; \beta_0) \right] \\ =& E\left[\frac{\partial f(Y_i|X_i; \beta_0)/\partial \beta}{f(Y_i|X_i; \beta_0)} \right] \\ =& \int \frac{\partial f(y|X_i; \beta_0)/\partial \beta}{f(y|X_i; \beta_0)} f(y|X_i; \beta_0) dy \\ =& \frac{\partial}{\partial \beta} \underbrace{\int f(y|X_i; \beta_0) dy}_{=1} \\ =& 0 \end{aligned}$$

Thus, an important condition of CLT is met: zero mean. By CLT,

$$n^{-1/2} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \log \ell_i(\beta_0) \to_d N(0, \Omega)$$

where the variance is a $k \times k$ matrix,

$$\Omega = E\left[\frac{\partial}{\partial \beta} \log \ell_i(\beta_0) \frac{\partial}{\partial \beta'} \log \ell_i(\beta_0)\right]$$

Combining these two results,

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d N(0, V)$$

where,

$$\begin{aligned} V =& \left(E\left[\frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_0)\right]\right)^{-1} E\left[\frac{\partial}{\partial \beta} \log \ell_i(\beta_0) \frac{\partial}{\partial \beta'} \log \ell_i(\beta_0)\right] \left(E\left[\frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_0)\right]\right)^{-1} \\ =& -\left(E\left[\frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_0)\right]\right)^{-1} \\ =& \left(J(\beta_0)\right)^{-1} \end{aligned}$$

We can show the second line as follows,

$$\begin{aligned} E\left[\frac{\partial^2}{\partial \beta \partial \beta'} \log \ell_i(\beta_0)\right] =& E\left[\frac{\partial}{\partial \beta'} \frac{\partial f(Y_i|X_i;\beta_0)/\partial \beta}{f(Y_i|X_i;\beta_0)}\right] \\ =& E\left[\frac{\partial}{\partial \beta'} \frac{\partial f(Y_i|X_i;\beta_0)/\partial \beta}{f(Y_i|X_i;\beta_0)}\right] \\ =& E\left[\frac{\partial^2 f(Y_i|X_i;\beta_0)/\partial \beta \partial \beta'}{f(Y_i|X_i;\beta_0)} - \frac{(\partial f(Y_i|X_i;\beta_0)/\partial \beta)(\partial f(Y_i|X_i;\beta_0)/\partial \beta')}{f(Y_i|X_i;\beta_0)^2}\right] \\ =& \underbrace{E\left[\frac{\partial^2 f(Y_i|X_i;\beta_0)/\partial \beta \partial \beta'}{f(Y_i|X_i;\beta_0)}\right]}_{=0} - E\left[\frac{\partial}{\partial \beta} \log f(Y_i|X_i;\beta_0) \frac{\partial}{\partial \beta'} \log f(Y_i|X_i;\beta_0)\right] \\ =& -E\left[\frac{\partial}{\partial \beta} \log \ell_i(\beta_0) \frac{\partial}{\partial \beta'} \log \ell_i(\beta_0)\right] \end{aligned}$$

where the first term is 0 by the same arguments used to demonstrate the zero-mean CLT condition. □

In this application, with a binary Y, the Hessian matrix is given by,

$$H_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \frac{f'(X_i'\beta)F(X_i'\beta) - f(X_i'\beta)^2}{F(X_i'\beta)^2} - (1-Y_i) \frac{f'(X_i'\beta)(1 - F(X_i'\beta)) + f(X_i'\beta)^2}{(1 - F(X_i'\beta))^2}\right] X_i X_i'$$

The solution is given by taking the derivative of the score function, applying the quotient rule.

For $F(\cdot) = \Phi(\cdot)$, you can show $f'(z) = -zf(z)$. Therefore, the Hessian matrix for a probit model is,

$$
\begin{aligned}
H_n(\beta) =& \frac{1}{n}\sum_{i=1}^{n}\left[Y_i\frac{-X_i'\beta\phi(X_i\beta)\Phi(X_i'\beta)-\phi(X_i'\beta)^2}{\Phi(X_i'\beta)^2}-(1-Y_i)\frac{-X_i'\beta\phi(X_i\beta)(1-\Phi(X_i'\beta))+\phi(X_i'\beta)^2}{\left(1-\Phi(X_i'\beta)\right)^2}\right]X_iX_i' \\
=& -\frac{1}{n}\sum_{i=1}^{n}\phi(X_i\beta)\left[Y_i\frac{X_i'\beta\Phi(X_i'\beta)+\phi(X_i'\beta)}{\Phi(X_i'\beta)^2}+(1-Y_i)\frac{-X_i'\beta(1-\Phi(X_i'\beta))+\phi(X_i'\beta)}{\left(1-\Phi(X_i'\beta)\right)^2}\right]X_iX_i'
\end{aligned}
$$

For a logit model, we can use the simplified version of the score function to solve the Hessian matrix.

$$
H_n(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\left[\lambda(X_i'\beta)\right]X_iX_i' = -\sum_{i=1}^{n}\left[\Lambda(X_i'\beta)(1-\Lambda(X_i'\beta))\right]X_iX_i'
$$

In both cases, the Hessian matrix is a negative-definite matrix given the strict concavity of distributions. However, a variance-covariance matrix must be positive-definite, which is why the Jacobian matrix is the negative of the Hessian.

We say that the approximate distribution of $\hat{\beta}^{ML}$ is,

$$
\hat{\beta}^{ML} \overset{a}{\sim} N(\beta_0, \hat{V}_n/n)
$$

where $\hat{V}_n = -H_n(\hat{\beta})^{-1}$ is the hessian matrix evaluated at $\hat{\beta}_n^{ML}$. Under certain regularity conditions, we know that $-H_n(\hat{\beta}) \to_p J(\beta_0)$ as $n \to \infty$ by WLLN, since $\hat{\beta}_n^{ML} \to_p \beta_0$ and the likelihood functions in question are continuously differentiable. This ensures that $\hat{V}_n \to_p V$. This is where the scaling of $1/n$ is important.

7 Interpretation & Fit

A general problem with the probit and logit models is the interpretability of the coefficients. From a latent variable model perspective, β has a clear marginal-effect interpretation (just as in a linear model). However, the latent variable is not a well-defined variable.

Moreover, the discrete outcome need not consistently vary with X_i. Suppose you change X_i by a value δ. Then,

$$
Y_i' - Y_i = \mathbf{1}\{X_i'\beta + \delta'\beta + \varepsilon_i \geq 0\} - \mathbf{1}\{X_i'\beta + \varepsilon_i \geq 0\}
$$

could take on values $\{-1, 0, 1\}$, depending on the error term. Thus, the marginal effect of X_i on Y_i is by definition heterogeneous.

7.1 Marginal Effects

For this reason, we focus on the effect of X_i on the $E[Y_i|X_i] = Pr(Y_i = 1|X_i)$. For a continuous regressor this is given by,

$$ME = \frac{\partial E[Y_i|X_i]}{\partial X_i} = \frac{\partial F(X_i'\beta)}{\partial X_i} = f(X_i'\beta)\beta$$

For the probit model, this is $\phi(X_i'\beta)X_i$, while for logit it is (X_i')(1- (X_i'))X_i.

For a discrete covariate, then you should evaluate the difference between predicted probabilities:

$$ME = E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0] = F(X_i'\beta + \gamma) - F(X_i'\beta)$$

where γ is the coefficient on the discrete dummy variable D.

The marginal effect depends on X_i, which means that we can evaluate it at different values of X_i. A common choice is ME at the Mean (MEM):

$$MEM_n = f(\bar{X}'\hat{\beta})\hat{\beta}$$

One issue with this is the non-representativity of a the mean covariate value. For example, if X includes a categorical variable (represented by dummy variables), the mean is an observation that does not exist in the data.

Alternatively, you can then compute the Average ME (AME).

$$AME_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i'\hat{\beta})\hat{\beta}$$

You can show that $AME_n \to_p E[f(X_i'\beta_0)\beta_0]$.

7.2 Odds ratios

An alternative approach to interpretation that is more commonly applied to the logit model, is the odds ratio: the probability that $Y = 1$ divided by the probability that $Y = 0$.

For the logit model, this given by,

$$\frac{\Lambda(X_i'\beta)}{1 - \Lambda(X_i'\beta)} = exp(X_i'\beta)$$

And the log odds-ratio is just $X_i'\beta$.

The ratio of two odds ratios can provide a useful interpretation interpretation. Consider, if you increase the j-th regressor by a single unit then the change in the index is,

$$\frac{exp(X_i'\beta + \beta_j)}{exp(X_i'\beta)} = exp(\beta_j)$$

A 1-unit change in X_j increases the odds-ratio by a factor of $exp(\beta_j)$. This interpretation is common in medical journals, which favour logit models and the presentation of odds ratios. In general, Economics journals tend to prefer probit models, with the estimation of marginal effects.

7.3 Goodness of fit

Given the non-linearity of probit and logit models, we cannot use R^2 to describe goodness of fit. There are two alternatives:

1. **Number of correct predictions.**

The predicted value of the model is the predicted probability $Y_i = 1$: a number $\in [0, 1]$. However, the outcome is discrete: $Y_i \in \{0, 1\}$. Pick a threshold - for example, $\hat{\rho} = 0.5$ - and assign the predicted outcome as follows:

$$\hat{Y}_i = \begin{cases} Y_i = 1 & \text{if} \quad \hat{\rho} \geq 0.5 \\ Y_i = 0 & \text{otherwise} \end{cases}$$

Test what share of the data is correctly predicted.

2. **McFadden's R^2**

McFadden's R^2 is is given by the ratio of the log-likelihood of the model.

$$\text{pseudo } R^2 = 1 - \frac{\log(L_n(\hat{\beta}_U))}{\log(L_n(\hat{\beta}_R))}$$

where $\ln(L_n(\hat{\beta}_U))$ is the log-likelihood computed for the unrestricted model (the model with all regressors) and $\ln(L_n(\hat{\beta}_R))$ the log-lihood from the restricted model (with just a constant).

7.4 Likelihood Ratio Test

This the ratio of log-likelihoods of two models can also be used to test hypotheses of the form,

$$H_0 : h(\beta_0) = 0$$

Here, $h(\cdot)$ is a function from $R^k \to R^q$. These need not be linear hypotheses. There exists two tests for hypotheses of this form. The first is the Wald statistic, which assumes that $h(\cdot)$ is a continuous function and we know the distribution of a continuous of $h(\hat{\beta}^{ML})$.

$$\text{Wald-stat} = nh(\hat{\beta}^{ML})' \left(\frac{\partial h(\hat{\beta}^{ML})}{\partial \beta'} \hat{V}^{-1} \frac{\partial h'(\hat{\beta}^{ML})}{\partial \beta} \right) h(\hat{\beta}^{ML}) \to_d \chi^2_q$$

However, the second option is more common. This is the simpler Likelihood Ratio test derived using a restricted and unrestricted estimator. The restricted estimator is given by,

$$\hat{\beta}_R = \underset{\beta \in B_R}{\arg\max}\, n^{-1} \log L_n(\beta)$$

where $B_R = \{\beta \in B : \ h(\beta) = 0\}$. The LR test statistic is then easily computed as,

$$\text{LR-stat} = 2 \times \left(\log\left(L_n(\hat{\beta}_U)\right) - \log\left(L_n(\hat{\beta}_R)\right) \right) \to_d \chi^2_q$$

These two tests are asymptotically equivalent, but the LR-test is far easier to compute.

References

Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge university press.

Verbeek, Marno. 2017. *A Guide to Modern Econometrics*. John Wiley & Sons.