Lecture 9: Binary Choice Models

March 6, 2025

Average Partial Effect

- The most common application of binary response models is when we are interested in "explaining" a binary outcome in terms of some explanatory variables. Thus, we are interested in a conditional probability.
- We treat data censoring problems later. For now, we focus on the first situation. So, y is a binary (zero-one) variable. For example, $y =$ employed or $y =$ arrested. Given a set of (exogenous) covariates $\mathbf{x}$, we are interested in

$$P(y = 1|\mathbf{x}) = p(\mathbf{x}),$$

which is called the *response probability*. It is the probability of a "success," that is, $y = 1$.

Average Partial Effect

- As in regression, we are interested in the partial effects of the x_j on $p(\mathbf{X})$. For continuous x_j, these are usually

$$\frac{\partial p(\mathbf{X})}{\partial x_j}$$

.

- For discrete x_j, look at changes in the response probability (usually holding other variables fixed). For example, if $x_k = \text{train}$ (job training indicator) and y is an employment indicator,

$$p(x_1, \ldots, x_{k-1}, 1) - p(x_1, \ldots, x_{K-1}, 0)$$

is the effect of job training on the employment probability, at given values for the other covariates.

Average Partial Effect

- In nonlinear models generally, and binary response models specifically, it is often useful to have a single number to summarize the relationship between $P(y = 1|x)$ and x_j. In a linear model that is simply the coefficient.
- Generally, we might report an estimated *average partial effect (APE)*. The APE for a continuous x_j is

$$E_{\mathbf{x}}\left[\frac{\partial p(\mathbf{x})}{\partial x_j}\right],$$

  which means we average the partial effect across the population distribution of $\mathbf{x}$. This is a weighted average of the partial effects at each outcome $\mathbf{x}$.
- Suppose x_k is a binary variable. Then its APE is

$$E_{x(k)}[p(\mathbf{x}_{(k)}, 1) - p(\mathbf{x}_{(k)}, 0)]$$

  where $\mathbf{x}(k)$ is the $1 \times K$ vector with x_k excluded.

Average Partial Effect

- Some simple, useful facts about Bernoulli (zero-one) random variables are

$$E(y|\mathbf{x}) = P(y = 1|\mathbf{x}) = p(\mathbf{x})$$

$$Var(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})]$$

- So a binary variable has natural heteroskedasticity except in the special case where $p(\mathbf{x})$ does not depend on $\mathbf{x}$.
- Unlike variables that take on more than two values, there is a necessary link between the mean and the variance. It is not possible for $E(y|\mathbf{x}) = p(\mathbf{x})$ while $Var(y|x) = p(x)[1 - p(x)]$. (If, say, y takes values in $\{0, 1, 2, \dots\}$, $Var(y|\mathbf{x})$ need not be related to $E(y|\mathbf{x})$, even though that is true for popular distributions such as the Poisson.)

LPM: Linear Probability Model

- The linear probability model (LPM) models the response probability as a function linear in parameters. Absorbing an intercept into x, if we take the model literally we are assuming

$$P(y = 1|\mathbf{x}) = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k = \mathbf{x}\beta.$$

  Because this is also $E(y|\mathbf{x})$, we can use OLS to consistently estimate β. In fact, if the conditional mean is truly $\mathbf{x}\beta$, the OLS estimator is unbiased.

- Because $Var(y|\mathbf{x}) = \mathbf{x}\beta(1 - \mathbf{x}\beta)$ – a rare case where we know the functional form of heteroskedasticity – inference for OLS should be made robust to heteroskedasticity. As we know, this is easy to do.

LPM: Linear Probability Model

- The LPM is simple to estimate and interpret. The often cited drawbacks of the LPM include
  1. Nothing guarantees the OLS fitted values, $\hat{y}_i = \mathbf{x}_i\hat{\beta}$, are in the unit interval. As these are estimates of the $p(\mathbf{x}_i)$, one might worry about estimated probabilities above one or negative. (In practice, this is a minor issue.)
  2. While we can use various functional forms in $\mathbf{x}$, it is difficult to impose, in a simple way, diminishing effects of the x_j on the $p(\mathbf{x})$. For example, if $\hat{\beta}_j > 0$, increasing x_j increases $p(\mathbf{x}) = \mathbf{x}\beta$ by β_j, no matter the values of x_j or the other elements of $\mathbf{x}$. Logically, the effect must diminish at some point.
- A leading reason for going from the LPM to nonlinear models of $p(\mathbf{x})$ is to allow the partial effects to vary across different $\mathbf{x}$.

Index Models

- A general index model has the form

$$P(y = 1|\mathbf{x}) = G(\mathbf{x}\beta)$$

for some $G : \mathbb{R} \to (0, 1)$. That is, $0 < G(\cdot) < 1$. In most cases, $G(\cdot)$ is actually a cumulative distribution function for a continuous random variable with density $g(\cdot)$. Then, $G(\cdot)$ is strictly increasing, and the estimates are easier to interpret.

- The leading cases are $G(z) = \Phi(z)$ (probit) and

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

(logit).

A Digression of MLE

- MLE is straightforward. The general log likelihood for random draw i is

$$l_i(\beta) = (1 - y_i) \log[1 - G(\mathbf{x}_i\beta)] + y_i \log[G(\mathbf{x}_i\beta)].$$

- **Example (Probit)**: Suppose y_i is a scalar binary response, so it takes just two values, zero and one. Let $\mathbf{x}_i$ be a $1 \times K$ vector with $x_{i1} = 1$ for simplicity. Suppose y_i is generated by a linear latent variable model:

$$\begin{aligned} y_i^* &= \mathbf{x}_i\theta + e_i \\ e_i|\mathbf{x}_i &\sim \text{Normal}(0, 1) \\ y_i &= 1 \text{ if } y_i^* > 0 \\ &= 0 \text{ if } y_i^* \leq 0. \end{aligned}$$

A Digression of MLE

- A useful shorthand is

$$y_i = 1\{y_i^* > 0\}$$

where $1[\cdot]$ is the indicator function, equal to one if the statement in brackets is true, zero otherwise.

- The data we observe are (x_i, y_i), and we are usually interesting in the effects of x_i on y_i.

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(y_i^* > 0|\mathbf{x}_i) = P(\mathbf{x}_i\beta + e_i > 0|\mathbf{x}_i) \\ &= P(e_i > -\mathbf{x}_i\theta|\mathbf{x}_i) = 1 - \Phi(-\mathbf{x}_i\beta) = \Phi(\mathbf{x}_i\beta) \end{aligned}$$

where $\Phi(z) = \int_{-\infty}^{z} \phi(v)dv$ is the standard normal cdf and

$$\phi(v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right)$$

is the standard normal pdf.

A Disgression of MLE

- We have now completely characterized the conditional distribution:

$$f(1|\mathbf{x};\theta) = \Phi(\mathbf{x}\theta)$$

$$f(0|\mathbf{x};\theta) = 1 - \Phi(\mathbf{x}\theta)$$

or

$$f(y|\mathbf{x};\theta) = [1 - \Phi(\mathbf{x}\theta)]^{(1-y)}[\Phi(\mathbf{x}\theta)]^{y}, \quad y = 0, 1$$

$$= 0 \text{ if } y \notin \{0, 1\}.$$

A Disgression of MLE

- EXAMPLE (Probit): The log likelihood for a random draw i is

$$\ell_i(\theta) = (1 - y_i) \log[1 - \Phi(\mathbf{x}_i\theta)] + y_i \log[\Phi(\mathbf{x}_i\theta)]$$

$$\nabla_\theta \ell_i(\theta) = -(1 - y_i)\mathbf{x}_i \frac{\phi(\mathbf{x}_i\theta)}{1 - \Phi(\mathbf{x}_i\theta)} + y_i\mathbf{x}_i \frac{\phi(\mathbf{x}_i\theta)}{\Phi(\mathbf{x}_i\theta)}$$

$$= \phi(\mathbf{x}_i\theta)\mathbf{x}_i \left[\frac{y_i - \Phi(\mathbf{x}_i\theta)}{\Phi(\mathbf{x}_i\theta)[1 - \Phi(\mathbf{x}_i\theta)]} \right]$$

- Therefore, the score is

$$s_i(\theta) = \phi(\mathbf{x}_i\theta)\mathbf{x}_i' \left[\frac{y_i - \Phi(\mathbf{x}_i\theta)}{\Phi(\mathbf{x}_i\theta)[1 - \Phi(\mathbf{x}_i\theta)]} \right]$$

- and

$$E[s_i(\theta_0)|\mathbf{x}_i] = \phi(\mathbf{x}_i\theta_0)\mathbf{x}_i \left[\frac{E[y_i|\mathbf{x}_i] - \Phi(\mathbf{x}_i\theta_0)}{\Phi(\mathbf{x}_i\theta_0)[1 - \Phi(\mathbf{x}_i\theta_0)]} \right] = 0$$

because $E[y_i|\mathbf{x}_i] = P(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i\theta_0)$.

A Disgression of MLE

- The Hessian has the form

$$H_i(\theta) = \nabla_\theta^2 s_i(\theta) = -\left[\frac{\phi(\mathbf{x}_i\theta)^2}{\Phi(\mathbf{x}_i\theta)[1-\Phi(\mathbf{x}_i\theta)]}\right]\mathbf{x}_i'\mathbf{x}_i + L(\mathbf{x}_i,\theta)[y_i - \Phi(\mathbf{x}_i\theta)]$$

  where $L(\mathbf{x}_i,\theta)$ is the Jacobian of

$$\frac{\phi(\mathbf{x}_i\theta)\mathbf{x}_i'}{\Phi(\mathbf{x}_i\theta)[1-\Phi(\mathbf{x}_i\theta)]}$$

  .

- Under correct specification, we can use

$$A(\mathbf{x}_i,\theta_0) = -E[H_i(\theta_0)|\mathbf{x}_i] = \left[\frac{\phi(\mathbf{x}_i\theta_0)^2}{\Phi(\mathbf{x}_i\theta_0)[1-\Phi(\mathbf{x}_i\theta_0)]}\right]\mathbf{x}_i'\mathbf{x}_i$$

- Then,

$$\hat{A} = N^{-1}\sum_{i=1}^{N}\left[\frac{\phi(\mathbf{x}_i\hat{\theta})^2}{\Phi(\mathbf{x}_i\hat{\theta})[1-\Phi(\mathbf{x}_i\hat{\theta})]}\right]\mathbf{x}_i'\mathbf{x}_i \xrightarrow{p} A_0.$$

A Disgression of MLE

- So the "usual" asymptotic variance estimator is

$$\left(\sum_{i=1}^{N} \left[\frac{\phi(\mathbf{x}_i\hat{\theta})^2}{\Phi(\mathbf{x}_i\hat{\theta})[1-\Phi(\mathbf{x}_i\hat{\theta})]} \right] \mathbf{x}_i'\mathbf{x}_i \right)^{-1}$$

  which is easily seen to be positive definite when the inverse exists.
- Testing multiple hypotheses about β (we drop the "o" subscript for simplicity) – usually joint exclusion restrictions – is most easily done with the Wald and LR statistics. The former is commonly used in canned packages (in Stata, it is computed with the "test" command), and the LR statistic is easily obtained because the value of the log likelihood is reported routinely.

MLE Testing

- Under correct specification of the entire distribution, we do not need a fully robust statistic. The LR statistic is

$$LR = 2(\ell_{ur} - \ell_r) = 2\left[\sum_{i=1}^{N} \ell_i(\hat{\theta}) - \sum_{i=1}^{N} \ell_i(\tilde{\theta})\right]$$

  where $\hat{\theta}$ is the unrestricted estimator and $\tilde{\theta}$ is the estimator with Q smooth restrictions imposed. Under H_0,

$$LR \xrightarrow{d} \chi^2_Q.$$

Some practical issues of Probit/Logit

- More interesting is: What do we do with the estimates? Let x_j be continuous. Then

$$\frac{\partial p(x)}{\partial x_j} = \beta_j g(\mathbf{x}\beta)$$

  and, because $g(z) > 0$ (assume it is a continuous density), β_j gives the direction of the partial effect. But its magnitude depends on $g(\mathbf{x}\beta)$.
- For probit, the largest value of the scale factor is about $.4 = g(0)$. For logit, it is .25.
- For two continuous covariates, the ratio of the coefficients give the ratio of the partial effects, independent of $\mathbf{x}$.

$$\frac{\frac{\partial p(\mathbf{x})}{\partial x_j}}{\frac{\partial p(\mathbf{x})}{\partial x_h}} = \frac{\beta_j g(\mathbf{x}\beta)}{\beta_h g(\mathbf{x}\beta)} = \frac{\beta_j}{\beta_h}.$$

Some practical issues of Probit/Logit

- Two common summary measures are the estimated PEAs and APEs. The estimated PEA for a continuous variable is

$$\widehat{PEA}_j = \hat{\beta}_j g(\mathbf{x}\hat{\beta})$$

- The APE has more appeal, as we are averaging partial effects for actual units:

$$\widehat{APE}_j = \hat{\beta}_j \left[N^{-1} \sum_{i=1}^{N} g(\mathbf{x}_i \hat{\beta}) \right]$$

- To use the delta method, must account for randomness in $\mathbf{x}_i$, too. Bootstrap makes that easy.
- It makes no sense to compare magnitudes of coefficients across probit, logit, LPM. Comparing APEs is preferred.

Goodness of Fit

- In addition to reporting coefficients, standard errors, partial effects, and their standard errors, some additional goodness-of-fit measures are sometimes reported.
- Define, for each i, a binary predictor

$$\tilde{y}_i = \begin{cases} 1 & \text{if } G(\mathbf{x}_i\hat{\beta}) \geq .5 \\ 0 & \text{if } G(\mathbf{x}_i\hat{\beta}) < .5 \end{cases}$$

- We make a correct prediction if $y_i = 0$ and $\tilde{y}_i = 0$ or $\tilde{y}_i = 1$ and $y_i = 1$. Let N_0 be the number of observations with $y_i = 0$ and N_1 the number with $y_i = 1$, so that $N = N_0 + N_1$.

Goodness of Fit

- We can compute the percent correctly predicted for each of the outcomes, and the overall percent correctly predicted. If N_{00} is the number of observations with $y_i = 0$ and $\tilde{y}_i = 0$ and N_{11} is the number of observations with $\tilde{y}_i = 1$ and $y_i = 1$, then the proportions correctly predicted are

$$q_0 = \frac{N_{00}}{N_0}, \quad q_1 = \frac{N_{11}}{N_1}.$$

- The overall proportion correctly predicted is

$$q = \frac{(N_{00} + N_{11})}{N} = \left(\frac{N_0}{N}\right) q_0 + \left(\frac{N_1}{N}\right) q_1,$$

  which is a weighted average of the two.

Pseudo R^2

- LL_k log likelihood with all variables
- LL_1 log likelihood with only a constant

$$0 > LL_k > LL_1 \text{ so } |LL_k| < |LL_1|$$

$$Pseudo\ R^2 = 1 - \frac{|LL_k|}{|LL_1|}$$

- Bounded between 0-1
- Not anything like an R^2 from a regression

Pseudo R^2

- Because the Kullback-Leibler information criterion is maximized for the true density, the values of the log likelihoods can be used to choose among different non-nested models. In practice, it might be difficult to choose between, say, logit and probit.(Often the differences are practically unimportant, although they can be when fitted values at the extreme tails are important.)

Endogeneity: Continuous EEV

- If we want to allow nonconstant partial effects, we need to turn to nonlinear models.
- With a single EEV (for simplicity), consider the model

$$y_1 = \mathbb{I}[a_1 y_2 + \mathbf{z}_1 \delta_1 + u_1 > 0]$$

$$u_1 | \mathbf{z} \sim \text{Normal}(0, 1)$$

Endogeneity

- The Rivers-Vuong (1988) approach is to make a homoskedastic-normal assumption on the reduced form for y_2:

$$y_2 = \mathbf{z}\delta_2 + v_2 = \mathbf{z}_1\delta_{21} + \mathbf{z}_2\delta_{22} + v_2, \quad \delta_{22} \neq 0$$

$$v_2|\mathbf{z} \sim \text{Normal}(0, \tau_2^2)$$

- Can relax normality in two-step methods. In fact, sufficient is

$$u_1 = \theta_1 v_2 + e_1$$

$$e_1|v_2, z \sim \text{Normal}(0, 1 - \theta_1^2\tau_2^2)$$

$$y_1 = \mathbb{I}[a_1 y_2 + \mathbf{z}_1\delta_1 + \theta_1 v_2 + e_1 > 0]$$

so that

$$P(y_1 = 1|v_2, \mathbf{z}) = \Phi(a_{\rho 1} y_2 + \mathbf{z}_1\delta_{\rho 1} + \theta_{\rho 1} v_2),$$

where each coefficient is multiplied by $(1 - \rho_1^2)^{-1/2}$ and

$$\rho_1 = \theta_1\tau_2 = \text{Corr}(v_2, u_1).$$

The scaled coefficients are identified because we effectively observe $v_2 = y_2 - \mathbf{z}\delta_2$.

Endogeneity

- The RV two-step approach is
  - **(i)** OLS of y_2 on $\mathbf{z}$, to obtain the residuals, $\hat{v}_2$.
  - **(ii)** Probit of y_1 on $\mathbf{z}_1, y_2, \hat{v}_2$ to estimate the scaled coefficients. A simple t test on $\hat{v}_2$ is valid to test $H_0 : \theta_1 = 0$.
- The original coefficients, which appear in the partial effects, are easily obtained from the set of two-step estimates:

$$\hat{\alpha}_1 = \frac{\hat{\alpha}_{\rho 1}}{\left(1 + \hat{\theta}_{\rho 1}^2 \hat{\tau}_2^2\right)^{1/2}}$$

Endogeneity

- If we make the stronger assumption

$$(u_1, v_2)|\mathbf{z} \sim \text{BivariateNormal}$$

with $\rho_1 = \text{Corr}(u_1, v_2)$, then we can proceed with MLE based on

$$f(y_1, y_2|\mathbf{z}) = f(y_1|y_2, \mathbf{z})f(y_2|\mathbf{z}).$$

- The distribution $f(y_2|\mathbf{z})$ is straightforward because it is homoskedastic normal with a linear conditional mean.
- For $f(y_1|y_2, \mathbf{z})$ we have, for example,

$$P(y_1 = 1|y_2, \mathbf{z}) = \Phi\left(\frac{a_1 y_2 + \mathbf{z}_1\delta_1 + (\rho_1/\tau_2)v_2}{(1-\rho_1^2)^{1/2}}\right)$$

and then $P(y_1 = 0|y_2, \mathbf{z})$ is immediate. Then, all parameters – $\alpha_1, \delta_1, \rho_1, \delta_2, \tau_2$ are estimated jointly by MLE conditional on $\mathbf{z}$.

- The Stata command is "ivprobit." The same sorts of goodness-of-fit measures and partial effects are available, of course. For APEs, still might want to bootstrap the standard errors, confidence intervals.