Binary Response Models

William Brasic

The University of Arizona

Linear Probability Model (LPM)

Definition 1: Linear Probability Model (LPM)

A Linear Probability Model (LPM) uses OLS to estimate a model where the outcome variable y_i is binary.

- Outcome could be
  - Whether an individual graduated college or not.
  - Whether a firm will hire an individual or not.

Linear Probability Model (LPM)

Property 1: Linear Probability Model (LPM)

When our outcome y_i is binary, under MLR Assumptions 1-4,

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = \boldsymbol{x}_i' \boldsymbol{\theta}$$

where the coefficients represent the change in the probability that $y_i = 1$.

Linear Probability Model (LPM)

Example 1: Linear Probability Model (LPM)

Suppose upon estimation we get

$$\widehat{y_i} = \boldsymbol{x}_i' \widehat{\boldsymbol{\theta}} = 0.2 - 0.05 x_{i1} + 0.3 x_{i2}.$$

1. For a unit increase is x_1, the predicted probability that $y = 1$ decreases by 0.05.

Linear Probability Model (LPM)

Example 1: Linear Probability Model (LPM)

Suppose upon estimation we get

$$\widehat{y_i} = \boldsymbol{x}_i'\widehat{\boldsymbol{\theta}} = 0.2 - 0.05x_{i1} + 0.3x_{i2}.$$

1. For a unit increase is x_1, the predicted probability that $y = 1$ decreases by 0.05.
2. When $x_1 = 1$ and $x_2 = 2$, the predicted probability that $y = 1$ is $0.2 - 0.05 + 0.3(2) = 0.75$.

Linear Probability Model (LPM)

Example 1: Linear Probability Model (LPM)

Suppose upon estimation we get

$$\widehat{y_i} = \boldsymbol{x}_i'\widehat{\boldsymbol{\theta}} = 0.2 - 0.05x_{i1} + 0.3x_{i2}.$$

1. For a unit increase is x_1, the predicted probability that $y = 1$ decreases by 0.05.
2. When $x_1 = 1$ and $x_2 = 2$, the predicted probability that $y = 1$ is $0.2 - 0.05 + 0.3(2) = 0.75$.
3. When $x_1 = 1$ and $x_2 = 3$, the predicted probability that $y = 1$ is $0.2 - 0.05 + 0.3(3) = 0.75 = 1.05.$ (huh?)

Linear Probability Model (LPM)

Property 2: Linear Probability Model (LPM)

Economists and statisticians typically stay away from the LPM because:

- Predicted probabilities are not bounded in the unit interval.
- When y_i is binary, $\mathbb{V}[u_i|\boldsymbol{x}_i] = \boldsymbol{x}_i'\boldsymbol{\theta}(1 - \boldsymbol{x}_i'\boldsymbol{\theta})$.
  - Not too big of a problem because of HCSE.

- How should we estimate a model when the outcome is binary then?

Maximum Likelihood Estimation (MLE)

Question 1

Throughout the semester we've used OLS which generates a nice closed-form solution. However, when the model is non-linear in parameters, we can't use OLS to solve for the parameter estimates. What estimation strategy should we use in this case then?

Answer to Question 1

Maximum Likelihood Estimation (MLE)!

Maximum Likelihood Estimation (MLE)

Definition 2: Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a method used for estimating the parameters of a statistical model. Given a set of observations, MLE seeks to find the parameter values that maximize the likelihood function, which measures the probability of the observed data under the model.

- Intuition is that we want to find the parameters that maximize the likelihood of observing the data we did observe as then we likely have parameters close to ones that truly generated our data!

MLE Fundamental Probability Property

Property 3: MLE Fundamental Probability Property

Following from basic probability rules,

$$f_{Y,X}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = f_{Y|X}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \cdot f_X(\boldsymbol{x}_i).$$

- Recall that $\mathbb{P}(Y = y, X = x) = \mathbb{P}(Y = y \mid X = x)\mathbb{P}(X = x)$. This is where the above comes from.
- $\boldsymbol{\theta}$ is the $(k+1) \times 1$ vector of true parameters we want to estimate.
- $\boldsymbol{x}_i$ assumed to be given exogenously and, thus, not depend on $\boldsymbol{\theta}$, but some other set of parameters we don't care about.

MLE Fundamental Probability Property

Property 4: MLE Fundamental Probability Property

Following from basic probability rules, if $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ represent a random sample where $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)$ (a specific data point), then the joint distribution is a product of the marginal distributions:

$$f_{Z_1,\ldots,Z_n}(\boldsymbol{z}_i, \ldots, \boldsymbol{z}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{z}_i; \boldsymbol{\theta}).$$

- Think about it has the probability of observing our sample.

Likelihood Function

Definition 3: Likelihood Function

Under random sampling, the likelihood function is defined as

$$l_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{z}_i; \boldsymbol{\theta}).$$

Log-Likelihood Function

Definition 4: Log-Likelihood Function

Under random sampling, the log-likelihood function is defined as

$$\begin{aligned}\mathscr{L}_n(\boldsymbol{\theta}) &= \ln\left(l_n(\boldsymbol{\theta})\right) \\ &= \ln\left(\prod_{i=1}^{n} f(\boldsymbol{z}_i;\boldsymbol{\theta})\right) \\ &= \sum_{i=1}^{n} \ln f(\boldsymbol{z}_i;\boldsymbol{\theta}).\end{aligned}$$

- We take the log because its easier to differentiate.
- Remember the log of a product is the sum of the logs!

Maximum Likelihood Estimator (MLE)

Definition 5: Maximum Likelihood Estimator (MLE)

Under random sampling, the maximum likelihood estimator (MLE) is defined as

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathscr{L}_n(\boldsymbol{\theta}).$$

- The MLE is the vector of parameters that maximizes are likelihood function!

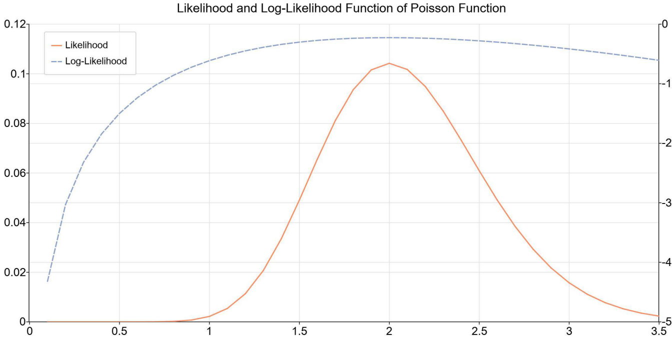
Maximum Likelihood Estimator (MLE)

Theorem 1: Maximum Likelihood Estimator (MLE)

Under random sampling, the maximum likelihood estimator (MLE) obtained by maximizing the likelihood function and the log-likelihood function are identical.

- Log of a function is a strictly increasing transformation which preserves the location of the maximum.

Maximum Likelihood Estimator (MLE)



Maximum Likelihood Estimator (MLE)

Property 5: Maximum Likelihood Estimator (MLE)

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln f(\boldsymbol{z}_i; \boldsymbol{\theta})$$

Maximum Likelihood Estimator (MLE)

Property 5: Maximum Likelihood Estimator (MLE)

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln f(\boldsymbol{z}_i;\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \ln\big(f_{Y|X}(y_i \mid \boldsymbol{x}_i;\boldsymbol{\theta}) \cdot f_X(\boldsymbol{x}_i)\big)\right]\end{aligned}$$

Maximum Likelihood Estimator (MLE)

Property 5: Maximum Likelihood Estimator (MLE)

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \ \sum_{i=1}^{n} \ln f(\boldsymbol{z}_i;\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \ \left[\sum_{i=1}^{n} \ln\left(f_{Y|X}(y_i \mid \boldsymbol{x}_i;\boldsymbol{\theta}) \cdot f_X(\boldsymbol{x}_i)\right)\right] \\
&= \arg\max_{\boldsymbol{\theta}} \ \left[\sum_{i=1}^{n} \ln\left(f_{Y|X}(y_i \mid \boldsymbol{x}_i;\boldsymbol{\theta})\right) + \sum_{i=1}^{n} \ln\left(f_X(\boldsymbol{x}_i)\right)\right]
\end{aligned}$$

Maximum Likelihood Estimator (MLE)

Property 5: Maximum Likelihood Estimator (MLE)

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \ \sum_{i=1}^{n} \ln f(\boldsymbol{z}_i; \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \ \left[\sum_{i=1}^{n} \ln \left(f_{Y|X}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \cdot f_X(\boldsymbol{x}_i) \right) \right] \\
&= \arg\max_{\boldsymbol{\theta}} \ \left[\sum_{i=1}^{n} \ln \left(f_{Y|X}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \right) + \sum_{i=1}^{n} \ln \left(f_X(\boldsymbol{x}_i) \right) \right] \\
&= \arg\max_{\boldsymbol{\theta}} \ \left[\sum_{i=1}^{n} \ln \left(f_{Y|X}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \right) \right].
\end{aligned}
$$

Maximum Likelihood Estimator (MLE)

Theorem 2: Maximum Likelihood Estimator (MLE) Properties

Under reasonable technical assumptions (beyond the scope of this class), the MLE is:

1. Consistent
2. Asymptotically normal
3. Asymptotically efficient

- All the properties (besides unbiasedness) we loved about OLS!

Maximum Likelihood Estimator (MLE)

Theorem 3: Maximum Likelihood Estimator (MLE) Properties

Under MLR Assumptions 1-6, the MLE for a linear regression model is identical to that produced by OLS.

- MLR Assumption 6 implies the distribution of the outcome given the covariates is normally distributed. Then, we set up the log-likelihood function using the normal PDF, take the first order condition with respect to $\boldsymbol{\theta}$, and solve for it getting $\widehat{\boldsymbol{\theta}} = (X'X)^{-1}X'\boldsymbol{y}$.

Maximum Likelihood Estimator (MLE)

Property 6: Pro and Con of MLE

1. Pro: We can now solve non-linear problems that don't have a nice closed form solution.
2. Con: Have to specify some functional form of the underlying data generating distribution.
   - Could lead to a misspecification issue and we don't get the estimates we truly want.

Latent and Observed Variable

Definition 6: Latent and Observed Variable

A latent variable $y_i^* = \boldsymbol{x}_i'\boldsymbol{\theta} + \epsilon_i$ is what we don't observe. In a binary response model (BRM), we observe $y_i = \mathbb{1}(y_i^* > 0)$.

- y_i^* could be wage and y_i could be whether or not the wage is above 50K.
- y_i is a binary variable.

Expectation of the Observed Outcome

Definition 7: Expectation of the Observed Outcome

Assuming ϵ_i and $\boldsymbol{x}_i$ are statistically independent and the CDF of the error ϵ_i is given by G, the expectation of the observed outcome is given by

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta})$$

Expectation of the Observed Outcome

Definition 7: Expectation of the Observed Outcome

Assuming ϵ_i and $\boldsymbol{x}_i$ are statistically independent and the CDF of the error ϵ_i is given by G, the expectation of the observed outcome is given by

$$\begin{aligned}\mathbb{E}[y_i \mid \boldsymbol{x}_i] &= f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) \\ &= \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i)\end{aligned}$$

Expectation of the Observed Outcome

Definition 7: Expectation of the Observed Outcome

Assuming ϵ_i and $\boldsymbol{x}_i$ are statistically independent and the CDF of the error ϵ_i is given by G, the expectation of the observed outcome is given by

$$\begin{aligned}\mathbb{E}[y_i \mid \boldsymbol{x}_i] &= f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) \\ &= \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) \\ &= \mathbb{P}(y_i^* > 0 \mid \boldsymbol{x}_i)\end{aligned}$$

Expectation of the Observed Outcome

Definition 7: Expectation of the Observed Outcome

Assuming ϵ_i and $\boldsymbol{x}_i$ are statistically independent and the CDF of the error ϵ_i is given by G, the expectation of the observed outcome is given by

$$\begin{aligned}\mathbb{E}[y_i \mid \boldsymbol{x}_i] &= f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) \\ &= \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) \\ &= \mathbb{P}(y_i^* > 0 \mid \boldsymbol{x}_i) \\ &= \mathbb{P}(\boldsymbol{x}_i'\boldsymbol{\theta} + \epsilon_i > 0 \mid \boldsymbol{x}_i)\end{aligned}$$

Expectation of the Observed Outcome

Definition 7: Expectation of the Observed Outcome

Assuming ϵ_i and $\boldsymbol{x}_i$ are statistically independent and the CDF of the error ϵ_i is given by G, the expectation of the observed outcome is given by

$$\begin{aligned}\mathbb{E}[y_i \mid \boldsymbol{x}_i] &= f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) \\ &= \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) \\ &= \mathbb{P}(y_i^* > 0 \mid \boldsymbol{x}_i) \\ &= \mathbb{P}(\boldsymbol{x}_i'\boldsymbol{\theta} + \epsilon_i > 0 \mid \boldsymbol{x}_i) \\ &= \mathbb{P}(\epsilon_i > -\boldsymbol{x}_i'\boldsymbol{\theta} \mid \boldsymbol{x}_i)\end{aligned}$$

Expectation of the Observed Outcome

Definition 7: Expectation of the Observed Outcome

Assuming ϵ_i and $\boldsymbol{x}_i$ are statistically independent and the CDF of the error ϵ_i is given by G, the expectation of the observed outcome is given by

$$\begin{aligned}
\mathbb{E}[y_i \mid \boldsymbol{x}_i] &= f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) \\
&= \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(y_i^* > 0 \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(\boldsymbol{x}_i'\boldsymbol{\theta} + \epsilon_i > 0 \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(\epsilon_i > -\boldsymbol{x}_i'\boldsymbol{\theta} \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(\epsilon_i \leq \boldsymbol{x}_i'\boldsymbol{\theta} \mid \boldsymbol{x}_i)
\end{aligned}$$

Expectation of the Observed Outcome

Definition 7: Expectation of the Observed Outcome

Assuming ϵ_i and $\boldsymbol{x}_i$ are statistically independent and the CDF of the error ϵ_i is given by G, the expectation of the observed outcome is given by

$$\begin{aligned}
\mathbb{E}[y_i \mid \boldsymbol{x}_i] &= f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) \\
&= \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(y_i^* > 0 \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(\boldsymbol{x}_i'\boldsymbol{\theta} + \epsilon_i > 0 \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(\epsilon_i > -\boldsymbol{x}_i'\boldsymbol{\theta} \mid \boldsymbol{x}_i) \\
&= \mathbb{P}(\epsilon_i \leq \boldsymbol{x}_i'\boldsymbol{\theta} \mid \boldsymbol{x}_i) \\
&= G(\boldsymbol{x}_i'\boldsymbol{\theta}).
\end{aligned}$$

PDF of Binary Variable

Definition 8: PDF of Binary Variable

- $f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) = G(\boldsymbol{x}_i'\boldsymbol{\theta})$.
- $f_{Y|X}(y_i = 0 \mid x_i; \boldsymbol{\theta}) = 1 - f_{Y|X}(y_i = 1 \mid x_i; \boldsymbol{\theta}) = 1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})$.

So, the conditional distribution of y_i given x_i is

$$f_{Y|X}(y_i \mid x_i; \boldsymbol{\theta}) = G(\boldsymbol{x}_i'\boldsymbol{\theta})^{y_i} \left[1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})\right]^{1-y_i}.$$

- Now we can set up our log-likelihood function for any G and use numerical methods (beyond the scope of this class) to solve for $\widehat{\boldsymbol{\theta}}$!
  - Newton's Method is one numerical solver.

ML for BRM

Definition 9: ML for BRM

The log-likelihood function for a BRM is given by

$$\begin{aligned}\mathscr{L}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln f_{Y|X}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \ln \left(G(\boldsymbol{x}_i'\boldsymbol{\theta})^{y_i} \left[1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})\right]^{1-y_i} \right)\end{aligned}$$

ML for BRM

Definition 9: ML for BRM

The log-likelihood function for a BRM is given by

$$
\begin{aligned}
\mathscr{L}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln f_{Y|X}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \ln \left(G(\boldsymbol{x}_i'\boldsymbol{\theta})^{y_i} \left[1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})\right]^{1-y_i} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[y_i \ln G(\boldsymbol{x}_i'\boldsymbol{\theta}) + (1 - y_i) \ln \left[1 - G(\boldsymbol{x}_i'\boldsymbol{\theta})\right] \right].
\end{aligned}
$$

Linear Probability Model

Definition 10: Linear Probability Model

When we assume G is the uniform CDF over the unit interval so $G(\boldsymbol{x}_i'\boldsymbol{\theta}) = \boldsymbol{x}_i'\boldsymbol{\theta}$, we have

$$\mathbb{E}[y_i \mid x_i] = G(\boldsymbol{x}_i'\boldsymbol{\theta}) = \boldsymbol{x}_i'\boldsymbol{\theta}.$$

- We get the LPM!
- Estimation by MLE or OLS will yield identical estimates of $\boldsymbol{\theta}$.

Probit Model

Definition 11: Probit Model

When we assume G is the standard normal CDF so $G = \Phi$ we get

$$\Phi(\boldsymbol{x}_i'\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}} \exp\left(\frac{-u^2}{2}\right) du,$$

the so-called probit model.

- Must lie within the unit interval alleviating concern of LPM!

Logit Model

Definition 12: Logit Model

When we assume G is the logistic CDF so $G = \Lambda$ we get

$$\Lambda(\boldsymbol{x}_i'\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\theta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\theta})},$$

the so-called logit model.

- Must lie within the unit interval alleviating concern of LPM!
- Often also called the logistic regression model.

Partial Effects

Definition 13: Partial Effects

Assuming G is either the logistic or standard normal CDFs, the partial effect of:

1. A continuous x_j on the probability that $y_i = 1$ is

$$\frac{\partial \mathbb{P}(y_i = 1 \mid \boldsymbol{x})}{\partial x_j} = \frac{\partial G(\boldsymbol{x}_i'\boldsymbol{\theta})}{\partial x_j} = g(\boldsymbol{x}_i'\boldsymbol{\theta})\theta_j.$$

Partial Effects

Definition 13: Partial Effects

Assuming G is either the logistic or standard normal CDFs, the partial effect of:

1. A continuous x_j on the probability that $y_i = 1$ is

$$\frac{\partial \mathbb{P}(y_i = 1 \mid \boldsymbol{x})}{\partial x_j} = \frac{\partial G(\boldsymbol{x}_i'\boldsymbol{\theta})}{\partial x_j} = g(\boldsymbol{x}_i'\boldsymbol{\theta})\theta_j.$$

2. A binary x_1 on the probability that $y_i = 1$ is

$$G(\theta_0 + \theta_1 + \theta_2 x_2 + \dots \theta_k x_k) - G(\theta_0 + \theta_2 x_2 + \dots \theta_k x_k).$$

Estimated Average Marginal Effect (AME)

Definition 14: Estimated Average Marginal Effect (AME)

The estimated average marginal effect (AME) of x_j on the probability that $y_i = 1$ is a summary measure of the partial effect of x_j given by

$$\widehat{\theta}_j \left[\frac{1}{n} \sum_{i=1}^{n} g\left(\boldsymbol{x}_i' \widehat{\boldsymbol{\theta}} \right) \right].$$

Single Hypothesis Tests

Definition 15: Single Hypothesis Tests

The test statistic for any hypothesis test when using the logit or probit models is the same as when using OLS with or without a binary outcome:

$$T = \frac{\widehat{\theta}_j - a}{\mathsf{se}\left[\theta_j\right]}$$

where a is the number we are testing θ_j against.

- Once forming the test statistic, we carry out single hypothesis tests as usual.

Multiple Hypothesis Tests

Definition 16: Multiple Hypothesis Tests

When testing multiple restrictions, two common test statistics are:

1. The Wald test statistic (complex formula beyond the scope of this class).
2. The likelihood ratio test statistic $LR = 2\left[\mathscr{L}_n\left(\widehat{\boldsymbol{\theta}}_{UR}\right) - \mathscr{L}_n\left(\widehat{\boldsymbol{\theta}}_R\right)\right]$.

- The LR statistic can be calculated by running probit/logit models for each of the two specifications.
- Both test statistics have an asymptotic χ_q^2 distribution where q is the number of restrictions.

Thank You!