Section 9: Probit, Logit, and RDD

Valentine Gilbert

November 11, 2021

Overview

1. Probit, Logit, and LPM
2. Maximum Likelihood Estimation
3. Regression Discontinuity

Probit, Logit, and LPM 1/6: Motivation

- Many outcomes of interest are binary:
  - Whether a student goes to college
  - Whether a parolee recidivates
  - Whether a baby is born prematurely
- OLS estimation of regressions with a binary outcome variable can give predicted probabilities that are less than 0 or greater than 1.
- Probit and logit regressions constrain predicted probabilities to be between 0 and 1

Probit, Logit, and LPM 2/6: In Defense of the LPM

- How can we justify using OLS for a binary outcome variable?
  - We're estimating the best linear approximation of the CEF, which isn't binary.
  - $E[Y_i|X_i] = P(Y_i = 1|X_i) = X_i\beta + u_i$
- If the regression is a saturated dummy-variable regression, the CEF *is* linear, so OLS gets us exactly what we want!
- "If the CEF is non-linear, regression approximates the CEF. Usually it does it pretty well. Obviously, the LPM won't give the true marginal effects from the right nonlinear model. But then, the same is true for the 'wrong' nonlinear model! The fact that we have a probit, a logit, and the LPM is just a statement to the fact that we don't know what the 'right' model is. Hence, there is a lot to be said for sticking to a linear regression function as compared to a fairly arbitrary choice of a non-linear one!" - Steve Pischke

Probit, Logit, and LPM 3/6: Probit and Logit Definition

- Probit and Logit simply use a CDF as a wrapper function to ensure that predicted values are bounded between 0 and 1.
- Probit uses the normal CDF:

$$P(Y_i = 1|X_i) = \Phi(X_i\beta)$$
$$\Phi(z) \equiv \sqrt{2\pi} \int_{-\infty}^{z} \exp(-x^2/2)dx$$

- Logit uses the logistic CDF:

$$P(Y_i = 1|X_i) = F(X_i\beta) = 1/(1 + \exp(-X_i\beta))$$

Probit, Logit, and LPM 4/6: Interpreting Probit and Logit

- Interpreting the coefficients from a probit or logit regression is difficult
- For a probit model, $\beta_1 = .5$ implies that a one unit increase in X_{1i} leads to an increase in the z-score of .5.
- How that translates into an increase in $P(Y_i = 1|X_i)$ depends on the value of $X_i\beta$.
- Similarly, for a logit model, $\beta_1 = .5$ implies that a one unit increase in X_{1i} leads to an increase in $P(Y_i = 1|X_i)$ of:

$$1/(1 + \exp(-X_i\beta - .5)) - 1/(1 + \exp(-X_i\beta))$$

- By contrast, in the LPM, $\beta_1 = .5$ is easy to interpret: A one unit increase in X_{1i} increases the probability that $Y_i = 1$ by .5.

Probit, Logit, and LPM 5/6: Interpreting Probit and Logit

- Because the marginal effect in the probit and logit models depends on the value of $X_i\beta$, translating results into probabilities requires choosing where to evaluate the marginal effects.
- One approach is to calculate average marginal effects:

$$\frac{1}{N}\sum_i \frac{\partial P(Y_i = 1|X_i)}{\partial X_{1i}}$$

- Another is to predict the marginal effect at a particular value of interest
  - e.g. the threshold in a regression discontinuity design

Probit, Logit, and LPM 6/6: Odds Interpretation

- The coefficients of a logit model can be interpreted in terms of the effect of an increase in a regressor on the odds that $Y_i = 1 | X_i$
- What are the odds? $P(Y_i = 1 | X_i) / P(Y_i = 0 | X_i)$
- If β_1 is the coefficient in a logit model, then e^{β_1} is the effect of a one-unit increase in X_{1i} on the odds.
- This is interpreted relative to 1:
  - $e^{\beta_1} = 1.1$ implies that a one-unit increase in X_{1i} increases the odds that $Y_i = 1$ by 10%
  - $e^{\beta_1} = .9$ implies that a one-unit increase in X_{1i} decreases the odds that $Y_i = 1$ by 10%
- You're asked to prove this result on the problem set.

MLE 1/5: Motivation

- We've talked about using probit and logit to estimate a model with a binary outcome variable
- We talked about when it makes sense to do so and about interpreting the coefficients of those models
- But how are those coefficients estimated?
- Probit and logit model parameters are typically estimated with maximum likelihood

MLE 2/5: OLS vs MLE

- So far in this class we've estimated our parameters by choosing estimates that minimize our mean squared prediction errors
- For example, OLS solves:

$$\min_{\beta} E\left[\left(Y_i - X_i'\beta\right)^2\right]$$

- Under this approach, we made no distributional assumptions about any of our random variables

MLE 3/5: OLS vs MLE

- Another approach is to choose the parameters that maximize the probability of observing the data you have
- That is, we solve:

$$\max_{\beta} f(Y|\beta)$$

- To solve this, we have to specify the density function, $f(x)$. MLE, therefore, relies on distributional assumptions, a potential disadvantage.
- But MLE has the advantage of being the most efficient estimator if our distributional assumptions are correct
- More importantly, MLE allows us to estimate nonlinear models, like logit and probit

MLE 4/5: Likelihood and log-likelihood

- Consider $f(Y_1 = y_1|\beta)$. Intuitively, this is the probability that you would observe the outcome you have in your data for the first observation conditional on β.
- Then what's the likelihood of observing your entire dataset conditional on β?

$$\begin{aligned}\mathcal{L}_N(\beta) &= f(Y_1 = y_1|\beta) \times f(Y_2 = y_2|\beta) \times \dots \times f(Y_N = y_N|\beta) \\ &= \prod_{i=1}^{N} f(Y_i = y_i|\beta)\end{aligned}$$

- What assumption did we make here?
- It's usually easier to work with the log-likelihood function:

$$L_N(\beta) = \sum_{i=1}^{N} \log f(Y_i = y_i|\beta)$$

MLE 5/5: Estimation

- Now that we have the log-likelihood function, how do we estimate $\hat{\beta}^{MLE}$?
- Maximize the likelihood by taking partial derivatives and solving the first-order conditions
- Suppose we're estimating a probit model. Then,

$$
\begin{aligned}
f(Y_i = y_i|\beta) &= \Phi(X_i\beta)^{y_i}(1-\Phi(X_i\beta))^{1-y_i} \\
\implies \log f(Y_i = y_i|\beta) &= y_i \log \Phi(X_i\beta) + (1-y_i)\log(1-\Phi(X_i\beta)) \\
\implies L_N(\beta) &= \sum_{i=1}^{N} y_i \log \Phi(X_i\beta) + (1-y_i)\log(1-\Phi(X_i\beta)) \\
\implies \frac{\partial}{\partial\beta}L(\beta) &= \sum_{i=1}^{N} y_i \frac{\phi(X_i\beta)X_i}{\Phi(X_i\beta)} - (1-y_i)\frac{\phi(X_i\beta)X_i}{1-\Phi(X_i\beta)} = \mathbf{0}_{k\times 1}
\end{aligned}
$$

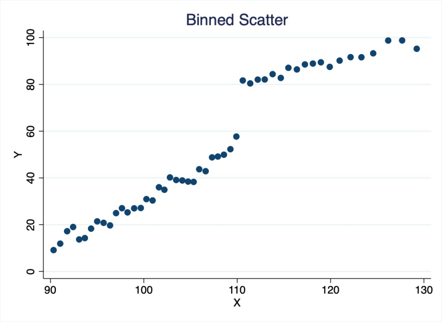
- Stata and R use numerical methods to solve this

Regression Discontinuity 1/3: Motivation

- Regression discontinuity design is a quasi-experimental method that exploits a cutoff in a continuous variable that determines eligibility for a treatment of interest
- Examples:
  - Students with combined SAT scores of less than 970 need a GPA of at least 3.0 to be admitted to Florida International University (Zimmerman JoLE 2014)
  - Workers seeking to organize need more than 50% of the vote to form a union (DiNardo and Lee QJE 2004)
  - Bombing targets in the Vietnam War were partly determined by an algorithm that calculated a continuous security score but presented output rounded to the nearest whole number (Dell and Querubin QJE 2017)
- In each case, the strategy is to identify a causal effect by comparing individuals just below the threshold to those just above the threshold.

Regression Discontinuity 2/3: Implementation

- Suppose you have a setting with an RD, leading to the following binned scatter plot:



- What regression specification estimates the size of the discontinuity at $X = 110$?
- $Y_i = \beta_0 + \beta_1(X_i - 110) + \beta_2 D_i + \beta_3(X_i - 110) \times D_i + u_i$, where D_i indicates $X_i \geq 110$.

Regression Discontinuity 3/3: Identifying Assumptions

- $Y_i = \beta_0 + \beta_1(X_i - 110) + \beta_2 D_i + \beta_3(X_i - 110) \times D_i + u_i$
- We're interested in a causal interpretation of β_2. Intuitively, under what conditions is that interpretation justified?
- If the only thing that changes discontinuously at the threshold is eligibility for the treatment in question, then the causal effect of that treatment is identified.
- When might this assumption fail?
  - If the threshold is one used by many different treatments (e.g. federal poverty guidelines)
  - If individuals sort to different sides of the threshold (e.g. if students can retake the SAT until they exceed an admissions cutoff)