Lecture 17: Binary Dependent Variables and Measurement Error

Pierre Biscaye

Fall 2022

Agenda

Binary dependent variables

- What if a qualitative variable is the *dependent* variable? For example
  - Whether a person completes high school
  - Whether a person is arrested
  - Whether a person is in the labor force
- We estimate $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$
- How do interpretations change?

Same assumptions still relevant

- If MLR1-MLR4 all hold then

$$E[y|x_1, x_2, ..., x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{1}$$

- Because y is binary

$$E[y|x_1, x_2, ..., x_k] = P(y = 1|x_1, x_2, ..., x_k) \tag{2}$$

$$P(y = 1|x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{3}$$

$$\frac{\Delta P(y = 1|x_1, ..., x_k)}{\Delta x_j} = \beta_j \tag{4}$$

Interpretation in words

- With a binary dependent variable, we interpret our estimates $\hat{\beta}_j$ as the *change in predicted probability* that $y = 1$ when x_j increases by one unit.
  - (Holding all of the other x variables constant.)
- We call this a *linear probability model* (LPM).
- Example: labor supply in Kenya during Covid-19.
  - Data from phone surveys for a representative panel of households.
  - Includes all household members age 18-64.
  - Model probability of paid employment in the past 7 days.

$$
\begin{aligned}
employed = &\beta_0 + \beta_1 age + \beta_2 gender + \beta_3 ishead + \\
&\beta_4 marital + \beta_5 currentnumadults + \\
&\beta_6 currentnum517 + \beta_7 currentnum04 + u
\end{aligned}
$$

To Jupyter!

Predicted values in the LPM

$$\hat{y}_i = p(\widehat{y_i = 1}|x) = \hat{\beta_0} + \hat{\beta_1} x_{i1} + \hat{\beta_2} x_{i2} + ... + \hat{\beta_k} x_{ik} \tag{5}$$

- For labor supply in Kenya during Covid-19, we estimated:

$$\begin{aligned} \widehat{employed} =& 0.219 - 0.001 age - 0.059 gender + 0.049 ishead - \\ & 0.003 marital - 0.019 currentnumadults - \\ & 0.006 currentnum517 + 0.003 currentnum04 \end{aligned}$$

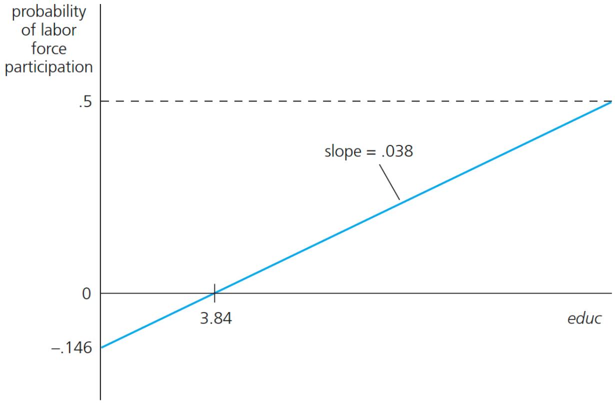
- What is the predicted probability of employment for a 30 year old woman who is not the head of household, is married, and lives in a multigenerational household with 4 total adults, 5 children aged 5-17, and 0 children aged 0-4?
- What if she is 60 years old?
- To Jupyter!

Predicted values, graphical example

Data on women's labor force participation in the US in 1975 (Wooldridge)

FIGURE 7.3

Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



Negative predicted probabilities in the LPM

- A negative predicted probability is nonsensical, as all probabilities must be between 0 and 1.
- But the linear probability model does not bound the range of predicted probabilities: $p(\widehat{y_i = 1}|x)$ could take any value
- We may or may not be concerned about this.
- Often we are more interested in marginal effects (our $\hat{\beta}$ estimates) and relatively less interested in predicted values.
- Predicted values outside the range of reason may not be relevant in sample
    - There is no individual in the Kenya sample with the characteristics we used for the prediction.
    - In the Wooldridge data, no one has less than 4 years of education.
- Still, important to be aware of. This is a bad feature of linear probability models.

Tradeoffs of using LPM

1. Predicted probabilities from regression aren't bounded between zero and one.
2. There must be heteroskedasticity in the linear probability model, since the variance of y—based on the probability $y = 1$—is now a function of our x variables. This violates our assumption of homoskedasticity:

   $$Var(u|x) = Var(u) = \sigma^2$$

   Therefore, our standard error calculations are more difficult
3. But, LPM coefficients are really easy to interpret and there are ways to deal with the above limitations.

Modeling binary dependent variables

- There are other models (logit, probit, tobit) used with binary dependent variables.
- These help deal with the limitations of LPM, but have their own disadvantages.
- LPM usually still useful for intuition.

Briefly: logit models

- Logistic regression, commonly referred to as *logit models* is specifically used to model binary dependent variables.
- The general mathematical equation for logistic regression is

$$y = 1/(1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots)}) \tag{6}$$

- This is called the inverse logistic function: hence the name logit model.
- A key characteristic is that the conditional distribution $y|x$ is a Bernoulli distribution rather than a Gaussian distribution (as in LPM), because the dependent variable is binary.
- Predicted values are probabilities and are therefore restricted to $[0, 1]$.
- Desmos graph example.

Interpretation with logit models

- Advantage: predicted values will all be between 0 and 1 - no impossible values.
- Disadvantage: interpretation is a little bit trickier than LPM.
- $\hat{\beta}_j$ estimates the change in *log odds* of the outcome for a one unit change in x_j.
- What are log odds?
  - Every probability can be expressed as the odds of being equal to 1. This is the ratio $p(y = 1)/p(y = 0) = p(y = 1)/(1 - p(y = 1)$.
  - Higher $p(y = 1)$ means greater *odds* of being equal to 1.
  - For example if $p(y = 1) = 0.8$, the odds that $p(y = 1)$ are $0.8/(1 - 0.8) = 0.8/0.2 = 4$. We would say the odds of being equal to 1 are 4 to 1.
  - *Log odds* are simply the natural log of the odds.

Probability and log odds

p	odds	logodds
.001	.001001	-6.906755
.01	.010101	-4.59512
.15	.1764706	-1.734601
.2	.25	-1.386294
.25	.3333333	-1.098612
.3	.4285714	-.8472978
.35	.5384616	-.6190392
.4	.6666667	-.4054651
.45	.8181818	-.2006707
.5	1	0
.55	1.222222	.2006707
.6	1.5	.4054651
.65	1.857143	.6190392
.7	2.333333	.8472978
.75	3	1.098612
.8	4	1.386294
.85	5.666667	1.734601
.9	9	2.197225
.999	999	6.906755
.9999	9999	9.21024

Why log odds?

We are modeling a probability p.

$$p = 1/(1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+...)}) \tag{7}$$

$$\frac{1}{p} = 1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+...)} \tag{8}$$

$$\frac{1-p}{p} = e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+...)} \tag{9}$$

$$\frac{p}{1-p} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+...)} \tag{10}$$

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... \tag{11}$$

$$\tag{12}$$

So you can see that β_j give changes in terms of the log odds that $y = 1$.

Logit example

$$employed = \beta_0 + \beta_1 gender + \beta_2 age + u$$

To Jupyter!

Logit example

$$employed = \beta_0 + \beta_1 gender + \beta_2 age + u$$

To Jupyter!

- $\beta_1 = -0.859$: being female decreases the log odds of being employed in the last 7 days by -0.859 relative to being male, holding age constant.
- $exp(-0.859) = 0.423$: the odds ratio of female to male is 0.423, holding age constant. Being female decreases the odds of employment by 57.7%.
- $\beta_2 = 0.015$: an additional year of age increase the log odds of being employed in the last 7 days by 0.015, holding gender constant.
- $exp(0.015) = 1.015$: the odds ratio for an additional year of age is 1.015. One year of age increase the odds of being employed by 1.5%.

Measuring the unmeasurable

- Suppose we wanted to estimate

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + u_i \tag{13}$$

- But we fear

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + \beta_3 Ability_i + e_i \tag{14}$$

- We could probably *never* measure ability in a satisfactory way. But omitting it will leave to omitted variable bias.
- If we can't directly measure "ability," what could we measure that relates to ability?

Proxy variables

- Suppose we measure IQ test scores.
- We consider using IQ as a proxy for ability.
- Suppose

$$Ability_i = \delta_0 + \delta_1 IQ_i + v_i \tag{15}$$

- What if we regress

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + \tilde{\beta}_3 IQ_i + \epsilon_i \tag{16}$$

How does this affect our model?

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + \beta_3 Ability_i + e_i \quad (17)$$

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + \beta_3(\delta_0 + \delta_1 IQ_i + v_i) + e_i \quad (18)$$

$$log(wage_i) = (\beta_0 + \beta_3\delta_0) + \beta_1 Ed_i + \beta_2 Exper_i \quad (19)$$
$$+\beta_3\delta_i IQ_i + (\beta_3 v_i + e_i)$$

- "New" MLR4: $E[\beta_3 v_i + e_i | Ed_i, Exper_i, IQ_i] = 0$.
- If it holds: $E[\hat{\beta_1}] = \beta_1$ and $E[\hat{\beta_2}] = \beta_2$.
- Using the proxy variable gives us *unbiased* estimators!

Unpacking the new error term

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + u_i \qquad (20)$$

$$log(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + \beta_3 Ability_i + e_i \qquad (21)$$

$$Ability_i = \delta_0 + \delta_1 IQ_i + v_i \qquad (22)$$

- We were worried about OVB in Eq 20: $E[Ability|Ed, Exp] \neq 0$.
- Assuming there are no omitted variables in Eq 21, we would have $E[u|Ed, Exp] = E[\beta_3 Ability + e|Ed, Exp] = \beta_3 E[Ability|Ed, Exp] \neq 0$.
- Controlling for *IQ* as a proxy for *Ability* gives a new error term $\epsilon_i = e_i + \beta_3 v_i$.
- MLR 4 now requires

$$E[\epsilon|Ed, Exp, IQ] = E[e + \beta_3 v|Ed, Exp, IQ] = \beta_3 E[v|Ed, Exp, IQ] = 0$$

- What we need (that is not already given) is $E[v|Ed, Exp] = 0$.

Proxy variables in words

- We started with a problem: Ability (which we can't measure) is likely correlated with Education.
- We proposed a solution: control for something which is correlated with Ability (IQ).
- If we control for IQ and *if* the part of Ability which is not explained by IQ (v) is uncorrelated with Education ($E[v|Ed, Exp] = 0$), then using this proxy variable will generate *unbiased* estimates.
  - We can think of a "good" proxy variable as one that captures all (or much of) the variation in the omitted variable that is correlated with the included X variables.
  - A "bad" proxy variable will not produce unbiased estimates, and might actually make you worse off.
- To Jupyter!

Measurement error

- One way to think about proxy variables: they are a version of the variable we care about that is measured with error.
  - IQ may be a (probably very) bad measurement of ability.
- Other variables may be measured with error, too. For example:
  - Wages may not be accurately reported.
  - Crime may be underreported
  - People may make mistakes in their education or age.
- How does meaurement error effect our estimates?

Example measurement error: reported age by round in Kenya household survey