POLS 6481. Research Design and Quantitative Methods II

Lecture 17. Binary Dependent Variables – Linear Probability Model

Readings: Wooldridge, *Introductory Econometrics 5e*, 7.5 + [7.7] + 8.5

Supplement: Aldrich and Nelson, *Linear Probability, Logit and Probit Models* 1–19 (or 1–30)

Spector and Mazzeo (1980) "Probit Analysis and Economic Education"

Outline

– Prelude: *'Personalized System of Instruction'*
1. Linear Probability Models
2. Out-of-Bounds Predictions
3. Heteroskedasticity & Weighted Least Squares
– Postlude: *'Personalized System of Instruction'* …

POLS 6481. Research Design and Quantitative Methods II

Lecture 17. Binary Dependent Variables – Linear Probability Model

Readings: Aldrich and Nelson, *Linear Probability, Logit and Probit Models* 1–19
Spector and Mazzeo (1980) "Probit Analysis and Economic Education"

– Prelude: *'Personalized System of Instruction'*

Probit Analysis and Economic Education
Spector, Lee C;Mazzeo, Michael
Journal of Economic Education (pre-1986); Spring 1980; 11, 2;
pg. 37

RESEARCH DESIGN. During the spring semesters of 1974 and 1975, Spector (1976) conducted experimental sections of principles of macroeconomics using the personalized system of instruction (PSI). In that experiment, using difference-of-means tests, he found that the PSI students scored significantly higher on the TUCE exam than a control group of students who took the course via the lecture method. He also found that the PSI students liked the PSI method, were interested in taking more courses of the PSI type, and would recommend a PSI course to others. Thus, this study indicated that PSI is a very effective method of teaching principles of macroeconomics.

In this study, Spector and Mazzeo reverse the direction of causality and instead examine how instructional method (PSI vs lecture) effected student performance in Spector's intermediate macroeconomics course, controlling for GPA and scores on the *Test of Understanding in College Economics* (*TUCE*) exam, as well as performance in principles of macro-economics and the student's major.

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Aldrich and Nelson, *Linear Probability, Logit and Probit Models* 1–19
Spector and Mazzeo (1980) “Probit Analysis and Economic Education”

– Prelude: *‘Personalized System of Instruction’*

Data: logist.dta
R Scripts: Lecture 17+18 Teaching.R

MACRO = intermediate macroeconomics grade (A = 4, E = 0)
GPA = cumulative grade point average
PRIN = principles-of-macroeconomics grade (A = 4, E = 0)
PSI = 1 if student took principles of macroeconomics via PSI (i.e., the student was in the original PSI group)
= 0 if the student did not (i.e., the student was in the original control group)

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Aldrich and Nelson, *Linear Probability, Logit and Probit Models* 1–19

– Prelude: *'Personalized System of Instruction'*

OBS PSI	GPA Grade	TUCE Grade	PSI	Grade	Letter Grade	OBS PSI	GPA Grade	TUCE Grade	PSI	Grade	Letter Grade
1	2.66	20	0	0	C	17	2.75	25	0	0	C
2	2.89	22	0	0	B	18	2.83	19	0	0	C
3	3.28	24	0	0	B	19	3.12	23	1	0	B
4	2.92	12	0	0	B	20	3.16	25	1	1	A
5	4.00	21	0	1	A	21	2.06	22	1	0	C
6	2.86	17	0	0	B	22	3.62	28	1	1	A
7	2.76	17	0	0	B	23	2.89	14	1	0	C
8	2.87	21	0	0	B	24	3.51	26	1	0	B
9	3.03	25	0	0	C	25	3.54	24	1	1	A
10	3.92	29	0	1	A	26	2.83	27	1	1	A
11	2.63	20	0	0	C	27	3.39	17	1	1	A
12	3.32	23	0	0	B	28	2.67	24	1	0	B
13	3.57	23	0	0	B	29	3.65	21	1	1	A
14	3.26	25	0	1	A	30	4.00	23	1	1	A
15	3.53	26	0	0	B	31	3.10	21	1	0	C
16	2.74	19	0	0	B	32	2.39	19	1	1	A

SOURCE: Spector and Mazzeo (1980).

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Aldrich and Nelson, *Linear Probability, Logit and Probit Models* 1–19

– Prelude: *'Personalized System of Instruction'*

OLS Model:

Dependent Variable: Grade MSE 0.150588 R-Square 0.4159

Variable	*DF*	*Parameter Estimate*	*Standard Error*	*T Ratio*	*Prob > |T|*
Intercept	1	–1.498	0.5239	–2.859	0.0079
GPA	1	0.4639	0.1620	2.864	0.0078
PSI	1	0.3786	0.1392	2.720	0.0111
TUCE	1	0.0105	0.0195	0.539	0.5944

WLS Model:

Weight: 1/(P(1-P)

Dependent Variable: Grade MSE 0.8121 R-Square 0.7622

Variable	*DF*	*Parameter Estimate*	*Standard Error*	*T Ratio*	*Prob > |T|*
Intercept	1	–1.309	0.2885	–4.536	0.0001
GPA	1	0.3982	0.08783	4.533	0.0001
PSI	1	0.3878	0.1052	3.687	0.0010
TUCE	1	0.0122	0.0045	2.676	0.0123

Later we'll discuss heteroscedasticity (which is an inherent problem; we'll discuss why too)

This will be the "football" pattern, for which, a few weeks ago, we discussed why OLS tends to yield standard errors that are too big (and so FGLS reduces standard errors and increases *t*'s)

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Wooldridge, *Introductory Econometrics 5e*, 7.5 + [7.7] + 8.5

1. Linear Probability Models

Wooldridge 7.5 $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

$$\Rightarrow \quad E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$$E(y|\mathbf{x}) = 1 \cdot P(y = 1|\mathbf{x}) + 0 \cdot P(y = 0|\mathbf{x})$$

$$\Rightarrow \quad P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$$\Rightarrow \quad \beta_j = \partial P(y = 1|\mathbf{x}) / \partial x_j$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Wooldridge, *Introductory Econometrics 5e*, 7.5 + [7.7] + 8.5

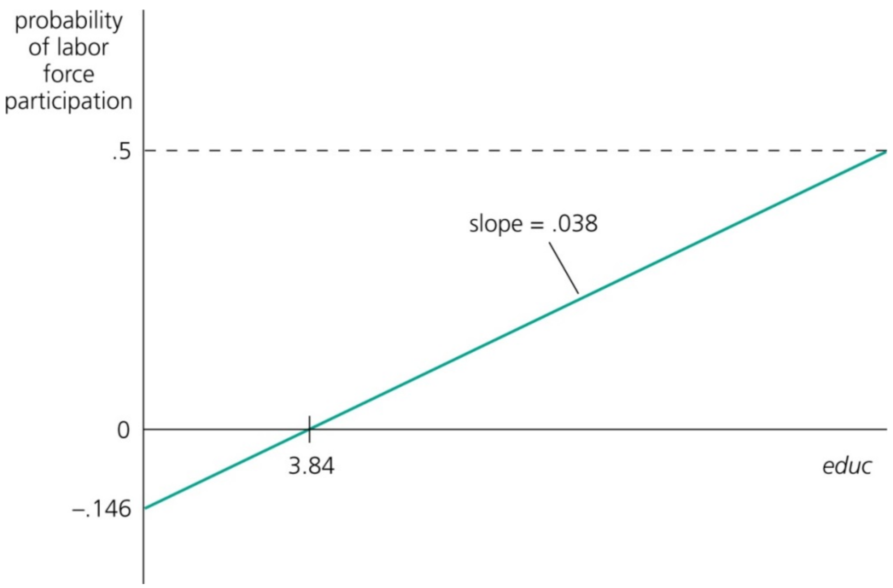
1. Linear Probability Models

Example (p. 249–251) Labor force participation for married women

$$\widehat{inlf} = \underset{(.154)}{.586} - \underset{(.0014)}{.0034}\, nwifeinc + \underset{(.007)}{.038}\, educ + \underset{(.006)}{.039}\, exper - \underset{(.00018)}{.00060}\, exper^2$$

$$- \underset{(.002)}{.016}\, age - \underset{(.034)}{.262}\, kidslt6 + \underset{(.0132)}{.0130}\, kidsge6$$

Graph predicted probabilities for *nwifeinc* = 50, *exper* = 5, *age* = 30, *kindslt6* = 1, *kidsge6* = 0; *educ* allowed to vary.



Maximum level of education in the sample is *educ* =17. Given the values of the other variables, this leads to a predicted probability to be in the labor force of about 50%.

Negative predicted probabilities for *educ* < 3.84, but no woman in the sample has *educ* < 5.

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Wooldridge, *Introductory Econometrics 5e*, 7.5 + [7.7] + 8.5

1. Linear Probability Models

Examples of Linear Probability Model:

- Example 7.12

R script: Lecture 17 Example7point12.R
Dataset: *CRIME1.DTA*

Sample is 2,725 young men in California, born during 1960 or 1961
Dependent variable is arrest during 1986 calendar year ($p = .277$)
Independent variables include proportion of prior arrests that led to conviction [*pcnv*], average sentence served in prior convictions [*avgsen*], months spent in prison [*tottime* & *ptime86*], number of quarters (0 to 4) of employment during 1986 [*qtime86*]

Ordinary least squares:

$$E(arr86) = .441 - .162*pcnv + .006*avgsen - .002*tottime - .022*ptime86 - .043*qemp86$$

POLS 6481. Research Design and Quantitative Methods II

Lecture 17. Binary Dependent Variables – Linear Probability Model

Readings: Wooldridge, *Introductory Econometrics 5e*, 7.5 + [7.7] + 8.5

1. Linear Probability Models

Advantages of Linear Probability Model:

- Easy estimation
- Easy interpretation
- In practice, estimated effects and predictions are often quite good

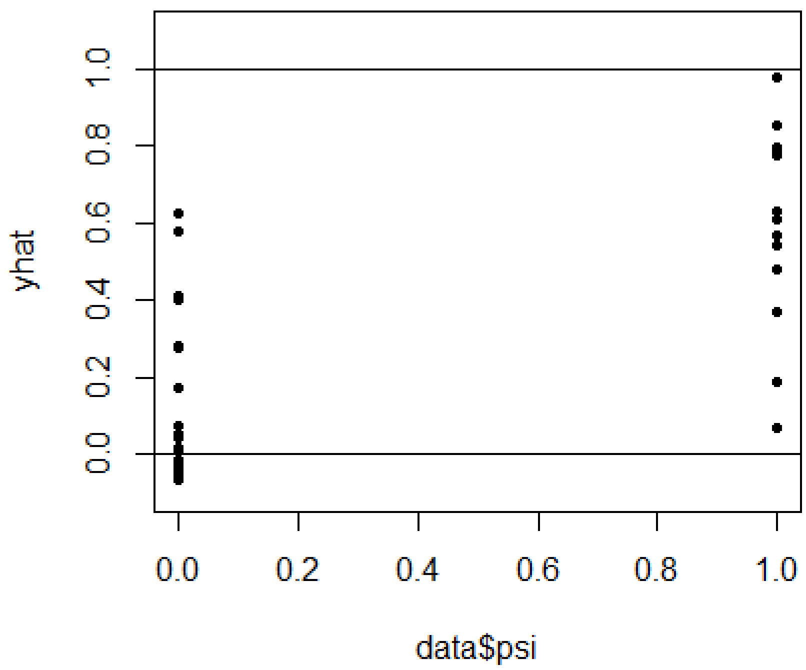
Disadvantages of Linear Probability Model:

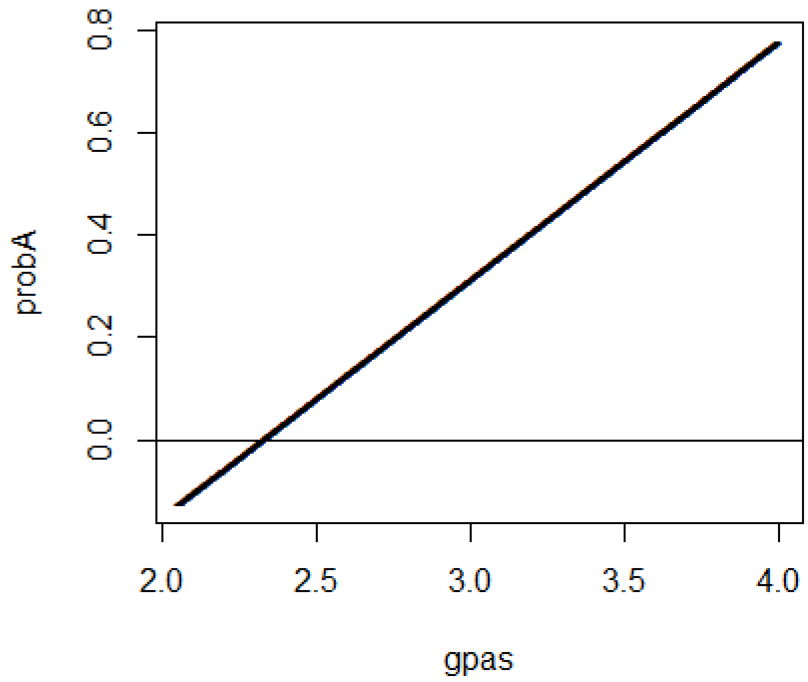
- Predicted probabilities may be outside unit interval (less than zero, greater than one)
- Heteroskedasticity

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Wooldridge, *Introductory Econometrics 5e*, 7.5 + [7.7] + 8.5

2. Out-of-Bounds Predictions

Because the fitted values are linear, predicted probabilities may be outside unit interval (less than zero, greater than one)





3. Heteroskedasticity

If Y_i takes on only two values, then u_i can assume only two values itself, for any given values of X_{ik}. That is, by equation 1.1 for Y_i equalling zero and one, respectively, we get:

$$\begin{aligned}&\text{If } Y_i = 0, \text{ then } (0 = \Sigma b_k X_{ik} + u_i) \text{ or } (u_i = -\Sigma b_k X_{ik})\\ &\text{If } Y_i = 1, \text{ then } (1 = \Sigma b_k X_{ik} + u_i) \text{ or } (u_i = 1 - \Sigma b_k X_{ik})\end{aligned} \qquad [1.7]$$

Now, it can be shown that the first key assumption about u_i, that its expectation be zero, can be maintained:

$$\begin{aligned}E(u_i) &= P(Y_i = 0)\,[-\Sigma b_k X_{ik}] + P(Y_i = 1)\,[1 - \Sigma b_k X_{ik}]\\ &= -[1 - P(Y_i = 1)]\,P(Y_i = 1) + P(Y_i = 1)\,[1 - P(Y_i = 1)] = 0\end{aligned}$$

As a result, OLS estimates of b_k will be unbiased.

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Aldrich and Nelson, *Linear Probability, Logit and Probit Models* 1–19

3. Heteroskedasticity

If Y_i takes on only two values, then u_i can assume only two values itself, for any given values of X_{ik}. That is, by equation 1.1 for Y_i equalling zero and one, respectively, we get:

$$\begin{aligned} &\text{If } Y_i = 0, \text{ then } (0 = \Sigma b_k X_{ik} + u_i) \text{ or } (u_i = -\Sigma b_k X_{ik}) \\ &\text{If } Y_i = 1, \text{ then } (1 = \Sigma b_k X_{ik} + u_i) \text{ or } (u_i = 1 - \Sigma b_k X_{ik}) \end{aligned} \qquad [1.7]$$

The assumption that the u_i have a constant variance cannot be maintained. In fact, the variance of u_i varies systematically with the values of the independent variables.

$$\begin{aligned} v(u_i) = E(u_i^2) &= P(Y_i = 0)[-\Sigma b_k X_{ik}]^2 + P(Y_i = 1)[1 - \Sigma b_k X_{ik}]^2 \\ &= [1 - P(Y_i = 1)][P(Y_i = 1)]^2 + P(Y_i = 1)[1 - P(Y_i = 1)]^2 \\ &= P(Y_i = 1)[1 - P(Y_i = 1)] \\ &= [\Sigma b_k X_{ik}][1 - \Sigma b_k X_{ik}] \end{aligned}$$

Thus… estimates of the sampling variances will not be correct, and any hypothesis tests or confidence intervals based on these sampling variances will be invalid.

3. Heteroskedasticity

Now consider the ith observation in our model. Since Y_i can only be 0 or 1, the ith disturbance term, ϵ_i, can take on only one of the following two values:

$$(1\text{-}5) \quad \epsilon_i = \begin{cases} 1 - X_i\beta & \text{if } Y_i = 1 \\ -X_i\beta & \text{if } Y_i = 0 \end{cases}$$

This result indicates that ϵ_i is not normally distributed and in fact it has a discrete distribution defined as:

Y_i	ϵ_i	$P(\epsilon_i)$	P
1	$1 - X_i\beta$	$-X_i\beta$	$1 - P$
0	$-X_i\beta$	$1 - X_i\beta$	P

For the estimates of β to be unbiased, the expected value of the disturbance term must equal 0. Therefore:

$$(1\text{-}6) \quad E(\epsilon_i) = P(-X_i\beta) + (1 - P)(1 - X_i\beta)$$
$$= (1 - X_i\beta)(-X_i\beta) + (X_i\beta)(1 - X_i\beta) = 0$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Spector and Mazzeo (1980) "Probit Analysis and Economic Education"

3. Heteroskedasticity

Now consider the ith observation in our model. Since Y_i can only be 0 or 1, the ith disturbance term, ϵ_i, can take on only one of the following two values:

$$(1\text{-}5) \quad \epsilon_i = \begin{cases} 1 - X_i\beta & \text{if } Y_i = 1 \\ -X_i\beta & \text{if } Y_i = 0 \end{cases}$$

This result indicates that ϵ_i is not normally distributed and in fact it has a discrete distribution defined as:

Y_i	ϵ_i	$P(\epsilon_i)$	P
1	$1 - X_i\beta$	$-X_i\beta$	$1 - P$
0	$-X_i\beta$	$1 - X_i\beta$	P

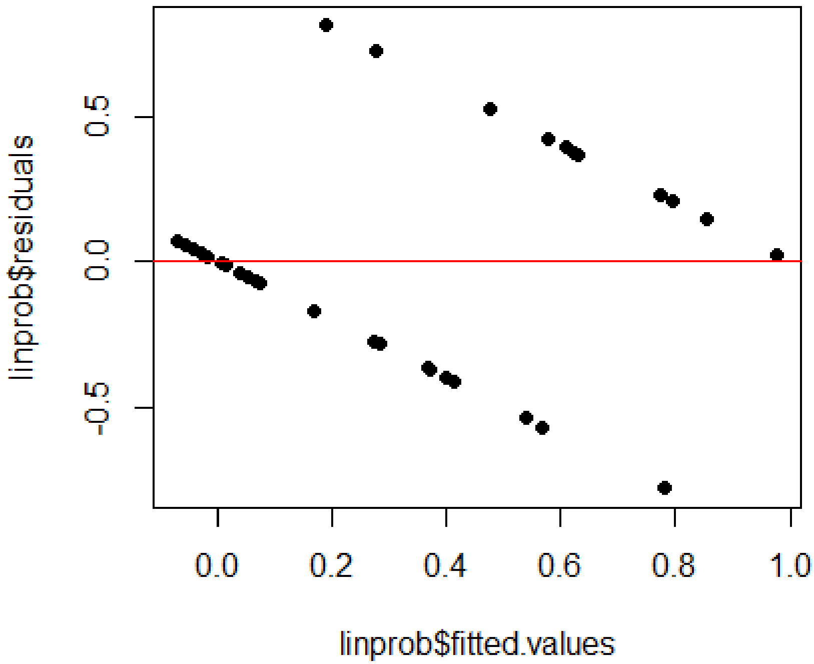
The variance of the disturbance term is equal to:

$$\begin{aligned} (1\text{-}7) \quad E(\epsilon_i)^2 &= (-X_i\beta)^2(1 - X_i\beta) + (1 - X_i\beta)^2(X_i\beta) \\ &= (X_i\beta)(1 - X_i\beta) \\ &= E(Y_i)[1 - E(Y_i)] \end{aligned}$$

Thus the variance is not constant, as assumed in (1-3). Rather, it depends on the values of the independent variables, and heteroskedasticity is present. As a result, the OLS estimators will no longer be the minimum variance linear estimators, and will yield incorrect results on t and F tests.

3. Heteroskedasticity

Algebra on last slide yields the following: $Var(y|\mathbf{x}) = P(y=1|\mathbf{x})\,[1 - P(y=1|\mathbf{x})]$



"goldbergerize" fitted values (6, 25–27): if $\hat{y}_i < 0$, re-set to .001; if $\hat{y}_i > 1$, re-set to .999

Weighted Least Squares (29–30): $\Rightarrow \;\; \hat{h}_i = \hat{y}_i(1 - \hat{y}_i)$

use inverse values of $\hat{h}_i$ as weights (automatic $\sqrt{\ }$)

POLS 6481. Research Design and Quantitative Methods II
Lecture 17. Binary Dependent Variables – Linear Probability Model
Readings: Wooldridge, *Introductory Econometrics 5e*, 7.5 + [7.7] + 8.5

3. Heteroskedasticity

Examples of Linear Probability Model:

- Example 8.9

R script: Lecture 17 Example8point9.R
Dataset: *GPA1.DTA*

Sample is 141 college students
Dependent variable is: computer ownership [*PC*]
Independent variables are: grades [*hsGPA*]
standardized test score [*ACT]*
dummy: at least one parent attended college [*parcoll*]

Ordinary least squares:

E(*PC*) = –.000 + .065**hsGPA* + .001**ACT* + .221**parcoll*

Weighted least squares (FGLS, with weights = $\frac{1}{\widehat{y_i}(1-\widehat{y_i})}$):

E(*PC*) = +.026 + .033**hsGPA* + .004**ACT* + .215**parcoll*