Linear Probability Model and Logistic Regression

Andy Grogan-Kaylor

28 Dec 2020 08:53:39

Introduction

The *Linear Probability Model* (LPM) is often discussed as an alternative to *logistic* regression. Essentially, the LPM is a linear model with a *dichotomous* dependent variable.

Setup

```
. clear all

. use http://www.stata-press.com/data/r15/margex, clear // artificial data from Stata
(Artificial data for margins)
```

Background

> I read through a number of references to develop this handout, especially the excellent book on Categorical Data Analysis by Long and Freese, and the always excellent Stata documentation. As I was finishing up this handout, I came across a superb handout by Richard Williams (referenced below), which does a better and more thorough job of explaining these issues than this short handout. You are encouraged to look it up.

Broadly speaking the Linear Probability Model is likely to give similar results to the logistic regression model: β coefficients are likely to have the same directions and similar statistical significances.

However, as one compares these approaches *more closely.* the Linear Probability Model is arguably incorrect on several grounds, some of which are illustrated in the figure below:

```
. twoway (lowess outcome age) (lfit outcome age), ///
> title("Outcome By Age") ///
> legend(order(1 "lowess smoother" 2 "linear fit")) ///
> scheme(michigan)

. graph export mygraph0.png, width(1000) replace
(file /Users/agrogan/Desktop/newstuff/categorical/LPM-and-logistic/mygraph0.png written in PN
> G format)
```

1. Marginal effects are mis-stated: The smoother indicates that the relationship of outcome and age is curvilinear. Thus, the effect on y of a 1 unit increase in x is different for different values of x.
2. Predictions can be implausible: By definition, negative probabilities are clearly impossible. However the linear fit predicts negative probabilities of the outcome for lower values of age.

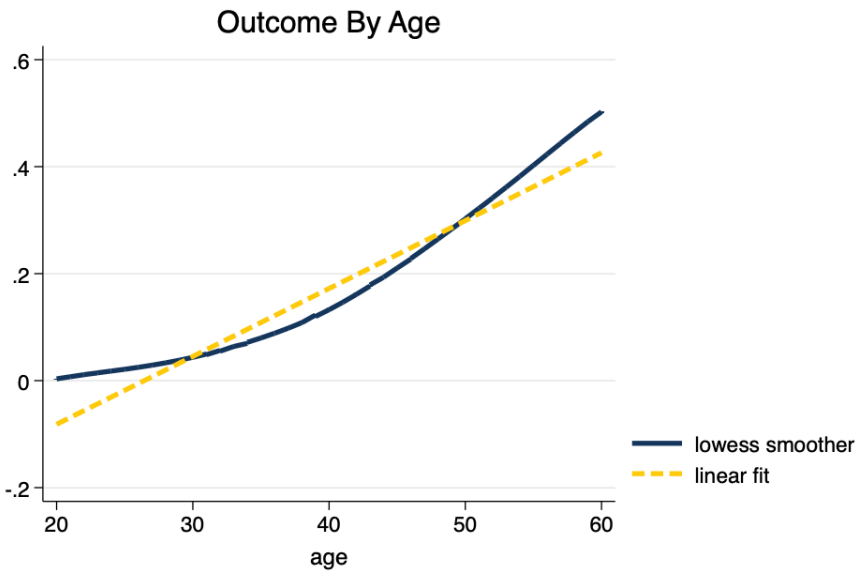

Figure 1: Lowess Smoother and Linear Fit Of Outcome By Age

3. Data with a dichotomous outcome are by definition heteroskedastic. The LPM (unless corrections are applied) makes assumptions of homoskedasticity. Thus, inferences about statistical significance–or the lack thereof–are likely to be incorrect.

   These differences in results are likely to become more salient the more one pays *detailed attention* to marginal effects for different values of the independent variables, and to predicted probabilities for different values of the independent variables.

Compare LPM and Logistic Regression In More Detail

Confirm That Outcome Is Dichotomous

```
. tabulate outcome // outcome is dichotomous
```

outcome	Freq.	Percent	Cum.
0	2,491	83.03	83.03
1	509	16.97	100.00
Total	3,000	100.00	

Linear Probability Model

```
. regress outcome sex##c.age i.group // linear probability model
```

Source	SS	df	MS	Number of obs	=	3,000
				F(5, 2994)	=	138.49
Model	79.386424	5	15.8772848	Prob > F	=	0.0000
Residual	343.253243	2,994	.114647042	R-squared	=	0.1878
				Adj R-squared	=	0.1865
Total	422.639667	2,999	.140926865	Root MSE	=	.3386

outcome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]

```
------------------------------------------------------------------------------
         sex |
      female |  -.2320346   .0489015    -4.74   0.000    -.3279185   -.1361508
         age |   .0061307   .0008814     6.96   0.000     .0044025    .0078589
             |
     sex#c.age |
      female |   .0072707   .0011613     6.26   0.000     .0049936    .0095477
             |
       group |
          2 |  -.0888273   .0164698    -5.39   0.000    -.1211206   -.0565339
          3 |  -.1034404   .0220694    -4.69   0.000    -.1467131   -.0601676
             |
       _cons |  -.0597978   .0401266    -1.49   0.136    -.1384763    .0188806
------------------------------------------------------------------------------
```

```
. predict yhat_LPM // predicted probabilities
(option xb assumed; fitted values)

. twoway scatter yhat_LPM age, ///
> title("Predicted Probabilities from Linear Probability Model") ///
> scheme(michigan)

. graph export myLPM.png, width(1000) replace
(file /Users/agrogan/Desktop/newstuff/categorical/LPM-and-logistic/myLPM.png written in PNG f
> ormat)
```

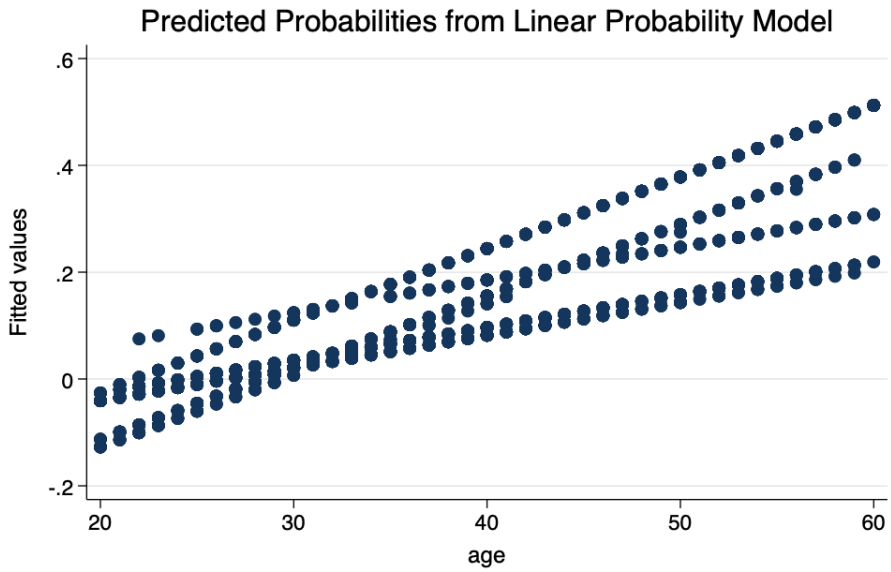

Figure 2: Predicted Values from Linear Probability Model

Logistic Regression

```
. logit outcome sex##c.age i.group // logistic regression model
Iteration 0:   log likelihood = -1366.0718
Iteration 1:   log likelihood =  -1118.129
Iteration 2:   log likelihood = -1070.8227
Iteration 3:   log likelihood = -1068.0102
Iteration 4:   log likelihood =   -1067.99
Iteration 5:   log likelihood =   -1067.99
Logistic regression                             Number of obs     =      3,000
                                                LR chi2(5)        =     596.16
```

```
                                                Prob > chi2       =     0.0000
Log likelihood =   -1067.99                     Pseudo R2         =     0.2182
```

outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex						
female	.5565025	.6488407	0.86	0.391	-.7152019	1.828207
age	.0910807	.0113215	8.04	0.000	.0688909	.1132704
sex#c.age						
female	-.001211	.0134012	-0.09	0.928	-.0274769	.025055
group						
2	-.5854237	.1349791	-4.34	0.000	-.8499779	-.3208696
3	-1.355227	.2965301	-4.57	0.000	-1.936416	-.7740391
_cons	-5.592272	.5583131	-10.02	0.000	-6.686545	-4.497998

```
. predict yhat_logistic // predicted probabilities
(option pr assumed; Pr(outcome))

. twoway scatter yhat_logistic age, ///
> title("Predicted Probabilities from Logistic Regression Model") ///
> scheme(michigan)

. graph export mylogistic.png, width(1000) replace
(file /Users/agrogan/Desktop/newstuff/categorical/LPM-and-logistic/mylogistic.png written in
> PNG format)
```

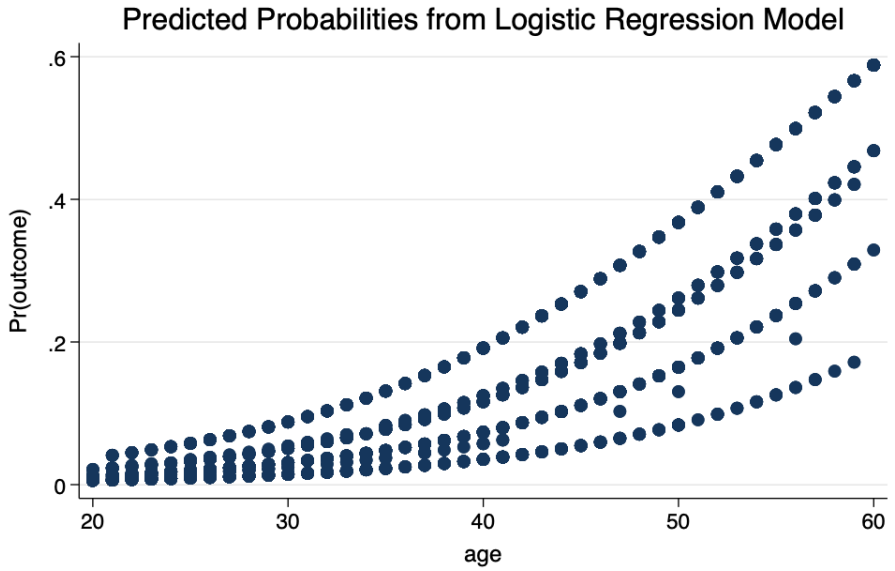

Figure 3: Predicted Values from Linear Probability Model

References

Long, J. S., & Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata (3rd ed.)*. College Station, TX: Stata Press.

StataCorp. 2019. *Stata 16 Base Reference Manual*. College Station, TX: Stata Press.

Williams, R. (2015). *Logistic Regression, Part I: Problems with the Linear Probability Model (LPM)*. South Bend, IN.