

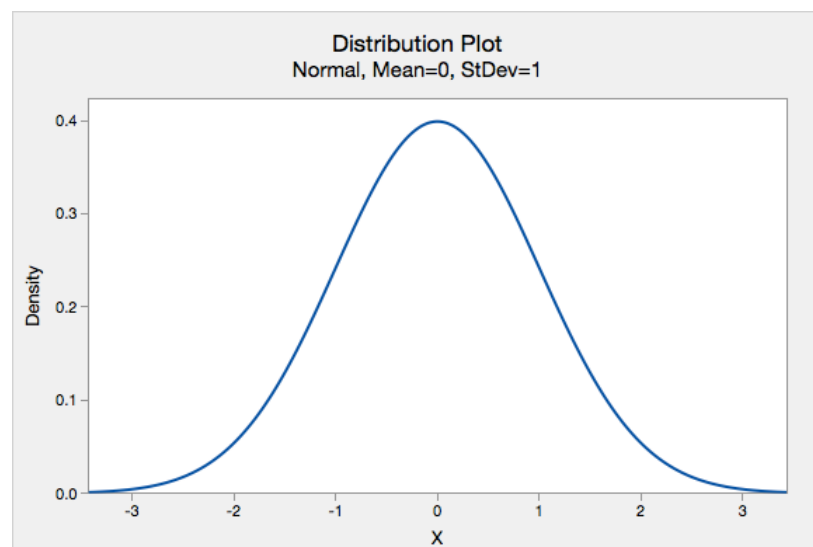
Topic 3 Simulation: Measurement error and the LPM

In the lectures this week we were concentrating on situations where the dependent variable is constrained to be either zero or one, an event either happens (1) or it does not (0). While we can include these variables on the right hand side of our regression equation with no problems, if they are the dependent variable things get a little more complex.

In class we discuss a few reasons why applying OLS (called a linear probability model when the dependent variable is dichotomous) will not be the best for this type of estimation. We mentioned that when using the LPM to create predicted probabilities it may produce estimates that lie outside the accepted range of 0 to 1. If our dependent variable is the probability of an event happening this is problematic. What would a negative figure represent? How do we interpret a value that is more than 100%? Using probits and logits allows us to avoid this issue. There are two other reasons the LPM is problematic that I want to quickly outline today, using some simulation to prove the latter of the two.

The first issue is what we would typically call “non-normality of the error term”. This means that we expect the error term to be normally distributed, looking something like the below Figure.

Figure 1: Normally Distribution



The normal distribution of the error term is not one of the Gauss Markov assumptions and thus is not needed for OLS to be BLUE. It does however matter when calculating p -values for significance testing..... when the sample size is very small. If we have a large enough sample size (more than 200 should be fine) the normality assumption is not really needed. However in small samples we need to be very careful with our inference and p -values.

Why is this normality assumption an issue when using a LPM? If we are concerned with modelling the likelihood of getting a degree (which is either 0 or 1 and our dependent variable) we might use the following set up where the likelihood of graduating for any individual (i) is based on the number of hours of study conducted (a simple but quite practical model!):

$$Degree_i = \beta_1 Hours\ of\ Study_i + \varepsilon_i$$

As Degree can only take two values here, we can work out the two possible values of ε_i .

When Degree=1 then we have:

$$1 = \beta_1 \text{Hours of Study}_i + \varepsilon_i$$

And we can rearrange to give:

$$1 - \beta_1 \text{Hours of Study}_i = \varepsilon_i$$

Likewise when Degree=0 then we have:

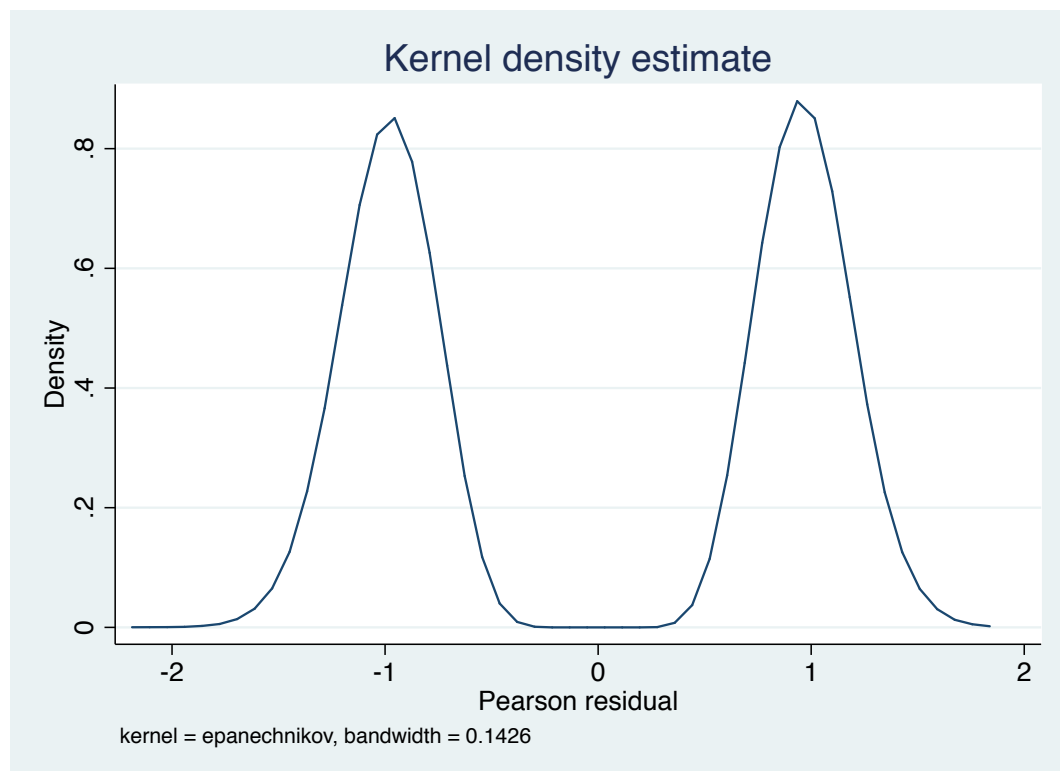
$$0 = \beta_1 \text{Hours of Study}_i + \varepsilon_i$$

And rearranging gives:

$$-\beta_1 \text{Hours of Study}_i = \varepsilon_i$$

Thus the only two values ε_i can take are $-\beta_1 X_i$ and $1 - \beta_1 X_i$. Thus our errors will be clustered at two points either side of zero, a very non-normal distribution. I have plotted the error term following a logit model below using a simulated dataset to try and give you a visual representation of what this may look like. Again, this is not really an issue unless the sample size is small but it is something worth being aware of.

Figure 2: Residual Plot following a Logit Regression



The second and more interesting (I believe anyway!) issue I want to raise with the LPM is how it performs when the dependent variable is misclassified. We have not, and will not aside from last weeks note, talked about measurement error in relation to regression models in this course. What do we mean by measurement error? We are talking about situations where our data fails to capture the true value of a variable.

Why does this happen? If someone asks you to fill in a form stating your income people may chose to slightly inflate the value that they record. Likewise if someone asks you to provide a value for the number of hours you have studied this week you are unlikely to have an accurate recording of this and may be forced to provide an estimate of the value. Thus there is some observation error. Measurement error associated with independent variables is usually a cause for concern in regressions and thus we should be careful to ensure that we are using reliable data based on good questionnaires. When we consider measurement error associated with the dependent variable this is usually less of a problem as the measurement error is absorbed into the error term and thus we just get slightly more variation in our estimation. LPM is slightly different though as we will try and demonstrate.

If our dependent variable can only be zero or one what possible measurement error can we have?

1. A value that should be coded as zero gets coded as one instead.
2. Or, a value that should be coded as one gets coded as zero.

This sort of measurement error leading to mis-classification of the data has been investigated by a number of authors. Notably, Hausman *et al.* (1998)¹ looked at the implications for the logit and probit models, as well as for the LPM. The negative implications of mis-classification are fundamentally worse for the LPM than they are for the probit and logit. There are issues associated both Logit and Probit models when the data are mis-classified, but these can be addressed. In the case of the LPM and measurement error than the parameters aren't identifiable. What does this mean? A model is identifiable if it is theoretically possible to learn the true values of this model's underlying parameters after obtaining an infinite number of observations from it. If it is not identifiable therefore we can never gain the true values of the underlying parameters and we are wasting our time!

What does this practically mean? When you have binary choice data, can you be absolutely sure that **every one of the observations** has been classified correctly into zeroes and ones? If your answer is "Yes", then you are probably lying to yourself or working with some very interesting data such as whether an individual is dead or not....! Which I guess could still be wrongly classified. If your answer is "No", then forget about using the LPM you will be trying to do the impossible, estimate parameters that aren't identified.

That's the technical part explained, can we show this on some simulated data? Let's start by creating an x variable, and an error term.

```
clear
set obs 10000
set seed 101
gen x = rnormal()
gen u = rnormal()
```

¹ The full reference is: [Hausman, J. A., J. Abrevaya & F. M. Scott-Morton](#), 1998. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87, 239-269.

Creating our dependent variable is a little trickier as we need a 0 or 1 variable, with the likelihood of being 1 depending on both x and u . The best way I could think of doing this is to create a linear version of y (lineary) as we have done in previous weeks, and then reclassify this linear version of y into a zero or one variable using the binomial distribution (rbinomial). This should give us the value of y where there is no measurement error and thus no misclassification. We can thus run the regression and see if we can find the estimates of the coefficient and x in an ideal situation.

```
gen lineary= .5 + .075*x + .075*u
gen y = rbinomial( 1, lineary )
reg y x
```

The output can be seen below in Figure 3 showing that the LPM pulls out good estimates of both x and the constant term.

Figure 3: Regression Out for LPM $y=0.5 + 0.75x$ when there is no measurement error

Source	SS	df	MS	Number of obs	=	10,000
Model	55.2791742	1	55.2791742	F(1, 9998)	=	226.07
Residual	2444.71923	9,998	.244520827	Prob > F	=	0.0000
				R-squared	=	0.0221
				Adj R-squared	=	0.0220
Total	2499.9984	9,999	.250024842	Root MSE	=	.49449

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0741902	.0049343	15.04	0.000	.064518	.0838623
_cons	.5002984	.0049451	101.17	0.000	.490605	.5099919

Let's now move onto looking at the same regression but when we have some measurement error in our dependent variable y . Let's create a new variable ($y2$) where 10% of our observations have the wrong y value (i.e. 1 when it should be 0 and vice versa). To do this I first created a variable that gives roughly 10% of 1 values. I then create a new y variable $y2$ which is equal to the original y value if the misclassified variable is equal to 0 (which should be 90% of the time). For the remaining 10% of values, when misclassified=1 we replace $y2$ to be the opposite of it's value in y . This would be a good point to browse the data and check that all has gone according to plan!

We can now run our regression this new version of the dependent variable and see how well our regression performs.

```
gen misclassified = rbinomial(1,.1)
gen y2= y if misclassified==0
replace y2 = mod(y+1,2) if misclassified==1
reg y2 x
```

What do the results in Figure 4 show us? We definitely have measurement error in Y . The estimates on x are biased towards zero, and already by quite a large factor given we should be finding a result of 0.075 and we are only finding a value that is four fifths of this 0.06. An interesting thing about this particular scenario is that we can easily identify when and how we

expect the bias to go. As the misclassification approaches half of the observations then we expect all of the coefficients (which are uncorrelated with the mis-classification) to also approach zero. If the unlikely case where to happen that the misclassification is apparent in more than half the cases then we would expect the signs on the coefficients to change to the opposite direction in the estimates.

Figure 4: Regression Output for LPM $y = 0.5 + 0.75x$ with measurement error

Source	SS	df	MS	Number of obs	=	10,000
Model	35.4250701	1	35.4250701	F(1, 9998)	=	143.72
Residual	2464.38133	9,998	.24648743	Prob > F	=	0.0000
				R-squared	=	0.0142
				Adj R-squared	=	0.0141
Total	2499.8064	9,999	.250005641	Root MSE	=	.49648

y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.059391	.0049541	11.99	0.000	.04968	.069102
_cons	.4961591	.004965	99.93	0.000	.4864268	.5058915

We can show how much of a difference the level of measurement error makes by increasing the measurement error from 10% of cases to 20% using the following code:

```
gen misclassified2 = rbinomial(1,.2)
drop y2
gen y2= y if misclassified2==0
replace y2 = mod(y+1,2) if misclassified2==1
reg y2 x
```

The estimate of x is smaller than it was, and by quite a large amount. This highlights that any measurement error is bad, and the degree of measurement error is also important in these types of models.

Figure 5: Regression Output for LPM $y = 0.5 + 0.75x$ with higher measurement error

Source	SS	df	MS	Number of obs	=	10,000
Model	23.6961486	1	23.6961486	F(1, 9998)	=	95.69
Residual	2475.89425	9,998	.247638953	Prob > F	=	0.0000
				R-squared	=	0.0095
				Adj R-squared	=	0.0094
Total	2499.5904	9,999	.249984038	Root MSE	=	.49763

y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0485741	.0049656	9.78	0.000	.0388404	.0583077
_cons	.4940573	.0049766	99.28	0.000	.4843022	.5038123

I guess the last thing we should do is to check whether the logit or probit performs better than the LPM. At this point in writing these notes I really hope they do perform better!!!

If I run a probit or a logit on this simulated data I see..... the marginal effects are about the same as the results provided by the LPM!

logit y2 x
margins, dydx(*)

This is rather upsetting! I looked at things a little further and the consensus view is that LPM does sometimes work, but sometimes it does not either. When simulating data I am aware of where the LPM has worked, but in real estimation I would not know the true value I am trying to find and thus I cannot safely use the LPM model. These problems are well documented by others including Horrace and Oaxaca (2006)² as well as a few keen econometricians such as: <http://davegiles.blogspot.co.uk/2011/09/econometrics-and-one-way-streets.html>.

The lesson learned from this is that there are many theoretical and empirical issues with the LPM. It may work as well as probits and logits some of the time, but if we know it is flawed in some circumstances we should avoid it. You do not know if you are using it in one of the cases where it will work.... or one of the cases where it will not! I know it is slightly easier to explain and to interpret than logits and probits but it should not be relied upon.

I hope some of the above was interesting, I am more than happy to discuss any of the factors we have explore today, I know there is a lot that you will not have seen before! Hopefully see you all next week for our next adventure into econometrics using simulation!

David Morris

² Full reference: Horrace, W. C. & R. L. Oaxaca, 2006. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90, 321-327.