Econometrics

Chapter 11: Regressions with Binary Dependent Variables

The Y variable is a 0/1, yes/no type variable.

Ex: does person get the job, do they get fired, are they employed, did they get the loan, etc.

Prediction:
Of course will never predict exactly a 0 or 1. So how to interpret the prediction - say of 0.8?
Well we would say an 80% chance of being unemployed. Or of those with those characteristics we expect 80% of them to be unemployed.
Why? First from the interpretation of the model as the CEF, then noting if the Y is 0-1 then the Expectation is also the probability.

Linear Probability Model

This is the model of simply using our linear OLS model in the setting where Y is 0-1.

And our β's are our change in probabilities from a unit change in X.

$Y_i = \beta_0 + \beta_1 X1 + \beta_2 X_2 ..... + \beta_k X_k + u_i$

$Pr(Y = 1|X_1, X_2 ....X_k) = \beta_0 + \beta_1 X1 + \beta_2 X_2 ..... + \beta_k X_k$

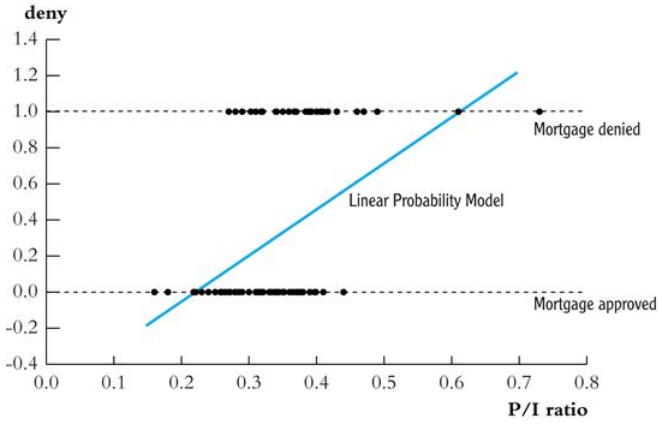

Figure 1: Linear Probability Model

All testing from before caries over (except R^2 - cannot ever be all on line unless all X's also binary and saturated).

But note that the error are always heteroskedastic in the LPM (think about it!). So be sure to hetero-robust SE's (though should always anyway).

Problems with the LPM:

Well because it is linear it implies that predicted probabilities can be greater than 1 and less than 0, which makes no sense.

Probit and Logit Regression

Both are nonlinear functions (in β's, before in X's) that force the predicted Y's between 0 and 1.

Both are based on CDF's - cumulative distribution functions.

Remember these are functions that give probability of some RV being less than some number and by definition is between 0 and 1

Probit - standard normal CDF
Logit - logistic CDF

Probit

$Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$

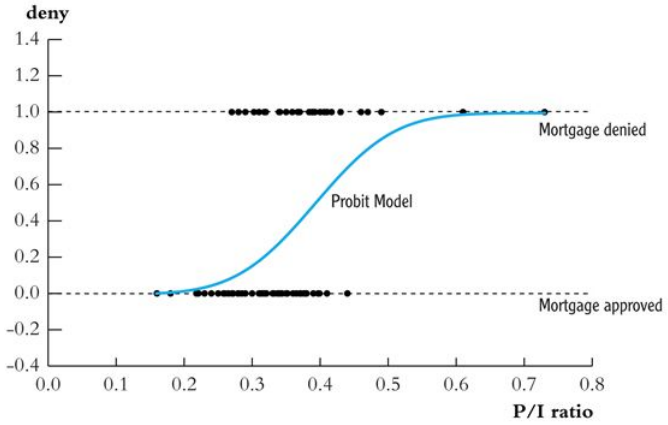

Figure 2: Probit Model

You should note that this graph looks just like a CDF - it is a CDF!

Say Y is whether a loan application is denied, and X is payment-to-income ratio, and $\beta_0 = -2$ and $\beta_1 = 3$.

What is the probability of a denial if P/I ratio is 0.4?

Well. $P(Y = 1|X) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8)$

Look up -0.8 in Cumulative normal in chart 1 and find $Pr(Z \leq -0.8) = 21.2\%$

So we expect about 21% of people with this P/I ratio to be denied and a random person with this P/I ratio to have 21% chance of being denied.

And note that the effect of P/I ration is not constant (linear) on Y. It depends where in the 'distribution' you are. But the 'z-value' value is linear.

Best to look up predicted probabilities before and after change in X to get interpretation.

The coefficients do not have a clear interpretation themselves except for their sign.

In case of one X can plot its effect on Y and will look just like a CDF with the S shape. Note will be flat around 1 and 0 and steep in middle.

Multiple X's in Probit

Just as in OLS need to worry about omitted variable bias, so should include all relevant and important X's

$Pr(Y = 1|X_1, X_2) = \Phi(\beta_0 + \beta_1 X + \beta_2 X_2)$

Note now the effect of changing X_1 depends on it and on X_2 because both effect where in 'distribution' you are (remember treat $\beta_0 + \beta_1 X + \beta_2 X_2$ like your Z (or t) that moves in the CDF of the standard normal)

Logit

Very similar to probit but instead of CDF of standard normal use that of the standard logistic function - denoted F().

$Pr(Y = 1|X_1, X_2...) = F(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k) = \frac{1}{1+exp[-(\beta_0+\beta_1 X_1+...+\beta_k X_k)]}$

As with probit for effects should look at changes because coefficients do not have direct interpretation.

Probit vs Logit vs LPM

In some cases LPM may do OK - especially if most of data is not near 0 or 1. But probably still do other because easy for software anyway

Probit vs logit? - in most cases really very identical in results. Used to use logit because of 'closed form solution' but with fast computers now not an issue. Most tend to default to probit.

Estimation

Probit coefficients are estimated with Maximum Likelihood (in other words not OLS).

Likelihood function: the joint probability distribution of the data, treated as a function of the unknown parameters.

Maximum Likelihood: picks values of the parameters that makes the Likelihood greatest (ie the probability of the data)

Forget the math part for a second and think about it in general terms.

Take the graph below. Assume we know the data came from a normal distribution $N(\mu, \sigma^2)$, and we know the variance of the distribution.

But we want to pick the mean. And we want to pick the mean so that the probability the data came from the distribution is as high as possible.

Which distribution would you pick? That is exactly what maximum likelihood is doing.

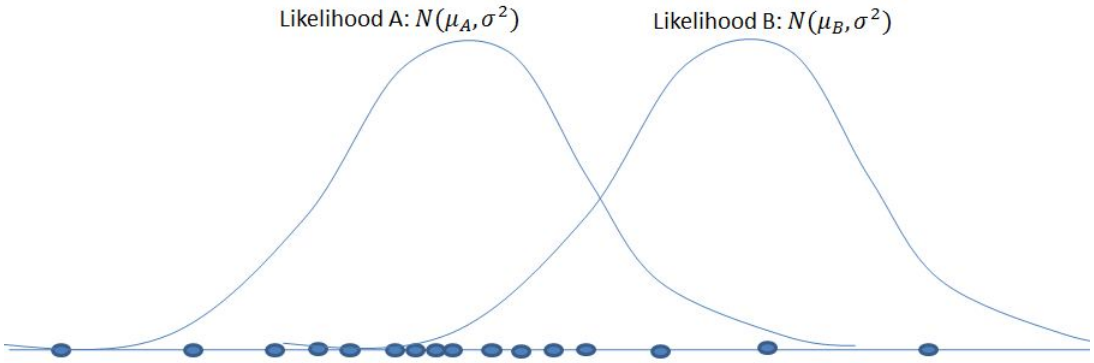

Figure 3: Maximum Likelihood

Of course while that gets the main idea, lets look at it closer.

Assume we have two i.i.d. draws from a Bernoulli distribution $Y = 0/1$.
So the only thing we don't know is p, ie. $Pr(Y = 1)$.

Well to pick the right p we first need to write down the joint probability
(recall the likelihood function is just the joint probability as a function of the unknown p).

So we want to write out $Pr(Y_1 = y_1, Y_2 = y_2)$.
But they are independent (i.i.d.) so:
$Pr(Y_1 = y_1, Y_2 = y_2) = Pr(Y_1 = y_1) \cdot Pr(Y_2 = y_2)$

Recall a Bernoulli distribution is $Pr(Y = y) = p^y(1-p)^{(1-y)}$.
If $y = 1 \rightarrow Pr(Y = 1) = p$ and $y = 0 \rightarrow Pr(Y = 0) = (1-p)$

So my joint probability of my data (my likelihood function) is: $[p^{y_1}(1-p)^{1-y_1}] \cdot [p^{y_2}(1-p)^{1-y_2}]$

But we write it as a function of the parameter, given the data (ie. likelihood *function*):
$f(p; y_1, y_2) = p^{(y_1+y_2)}(1-p)^{(2-(y_1+y_2))}$

So now just choose p to maximize $f(p; y_1, y_2)$.

In practice we maximize the *log* of the likelihood. Why?
Recall the ln function changes multiplication to addition, and changes powers to multiplication. This makes the math easier.

$\max_p ln(p^{(y_1+y_2)}(1-p^{(2-(y_1+y_2))}) = \max_p (y_1+y_2)ln(p) + (2-(y_1+y_2))ln(y-p)$

This is easily shown to give: $\hat{p} = \frac{y_1+y_2}{2}$

And in the general case ML yields $\hat{p} = \overline{y}$

Now this seems intuitive, and so is good to understand. The ML estimator of p (the probability of $Y = 1$), is just the sample average of $Y = 1$.

This same procedure extends to probit and logit (just the math is a bit messier because the probability function of a normal distribution is a bit more complicated that that for a Bernoulli).

And usually 'numeric' methods are actually used to find the solution - ie. some 'search' function.

MLE (maximum likelihood estimator) is consistent and normally distributed in large samples. So all testing etc. proceeds just as before.

Measures of Fit for MLE

1. Fraction correctly predicted. If $Y_1 = 1$ and predicted probability is $> 50\%$ or $Y_1 = 0$ and predicted probability is $< 50\%$ it is 'correctly predicted'.

This makes sense, but treats predicted probability of 51% and 95% the same.

2. pseudo-R^2: compares the value of the likelihood function you estimated to the value of the likelihood function with no regressors

But what is a 'good' measure of 'fit' for MLE is not really universally accepted.