

Intermediate Social Statistics Hilary 2009 Lecture : Binary Discrete Dependent Variable

Raymond Duch

Nuffield College

February 4, 2009

Continuous Dependent Variables

- Up until now, we have been treating all of our models as if Y were continuous.
- Today we'll consider the class of models where Y is non-continuous.
- Examples of continuous Y might include:
 - ▶ Presidential approval rates
 - ▶ Policy mood
 - ▶ Congressional polarization
 - ▶ Political tolerance
 - ▶ International trade
 - ▶ Globalization
 - ▶ Others?

Discrete Dependent Variables

- Lots of dependent variables cannot be characterized as continuous
- Those fall into several categories, such as (with examples):
 - ▶ Count (terrorist bombings)
 - ▶ Binary (votes)
 - ▶ Ordered (agree-to-disagree scales)
 - ▶ Multinomial (candidates in a primary; parties in multiparty election)
- And we'll treat these separately.

Functional Form of Discrete Models

All of our models will resemble probability models, of the sort:

$$\text{Prob}(\text{event } j \text{ occurs}) = \text{Prob}(Y=j) = F[\text{stochastic component, systematic component}]$$

Illustrations of Bivariate Dependent Variable

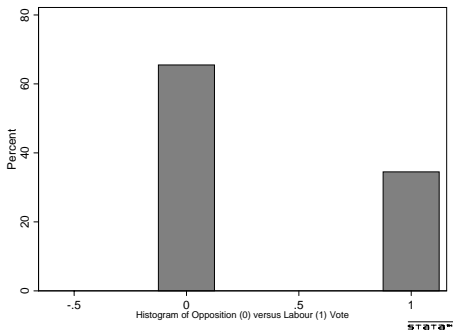
- This either occurs when the situation is genuinely binary—e.g., vote Labour or vote Opposition
- Or when the situation is continuous in the underlying (but unobserved) reality, but binary in observation—e.g., the decision to make, or not make, campaign contributions, which in a latent sense is a (continuous) probability model, but all we observe is [contribute, do not vote contribute].

An Example: Vote Preference of UK Citizens

- The example we will focus on in this lecture is from a 2004 survey of the voting preferences of U.K. citizens.
- The binary choice is vote Labour or vote for one of the opposition parties.

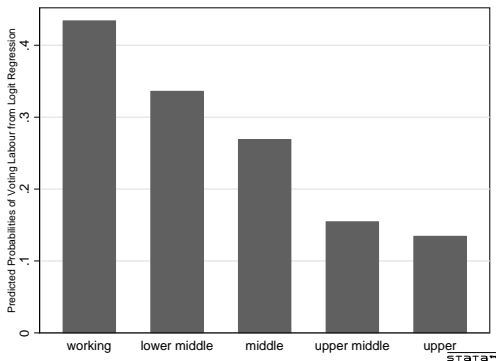
Illustrations of Bivariate Dependent Variable

Figure: Frequency of Labour versus Opposition Vote: UK 2004



Insights from Limited-Dependent Variable Models

Figure: Predicting Vote Choice Based on Class: UK 2004



What's wrong with the linear probability model?

- Why not just estimate:

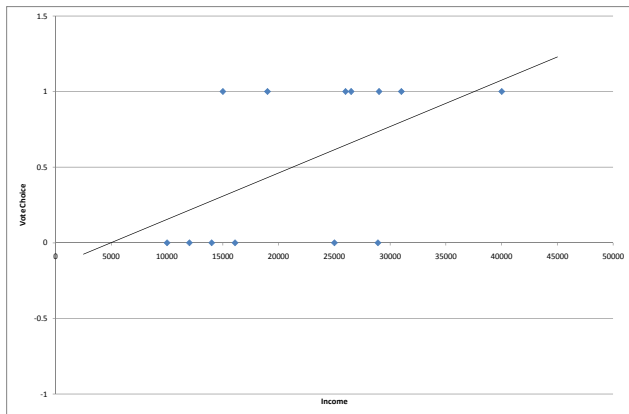
$$y = \mathbf{x}\beta + \varepsilon$$

- where $y = 0$ or $y = 1$?
- In terms of our example from the 2004 European Election study this would suggest,

$$\text{Labour Vote} = \text{Irsel} + \varepsilon$$

What's wrong with the linear probability model?

Figure: Estimating Hypothetical Vote Choice Model with OLS Regression



What's wrong with the linear probability model?

- First, you can see where ε will be heteroskedastic.
- The variance of it will be lowest around $p = 0.5$, and highest close to 0 and 1.
- But we can fix this with GLS. So this isn't too too too serious.

What's wrong with the linear probability model?

- Much more seriously, the model—you can see why—will make nonsense predictions, with $p < 0$ and $p > 1$.
- That will also produce negative variances. We can see this more clearly by estimating the following model using OLS:
- Labour vote = retnat + class + union + southwest + urban + lrsel
+ own + ε

What's wrong with the linear probability model?

Figure: Stata OLS Estimation of Labour Vote Model

Sunday January 27 06:45:29 2008 Page 1



```
1 . regress incumvote retnat class union southwest urban lrsel self own
```

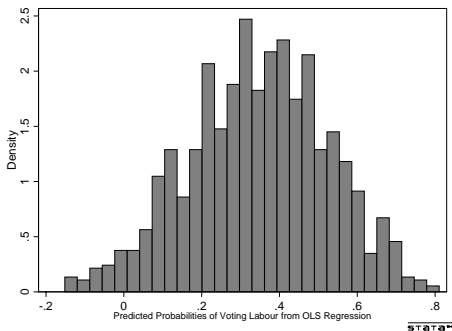
Source	SS	df	MS			
Model	26.5636924	7	3.79481321	Number of obs =	785	
Residual	151.798091	777	.195364338	F(7, 777) =	19.42	
Total	178.361783	784	.227502275	Prob > F =	0.0000	
				R-squared =	0.1489	
				Adj R-squared =	0.1413	
				Root MSE =	.442	

incumvote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
retnat	-.1366159	.0193703	-7.05	0.000	-.1746402	-.0985916
class	-.0713708	.0161261	-4.43	0.000	-.1030267	-.0397149
union	.0950899	.0383441	2.48	0.013	.0198196	.1703603
southwest	-.1541968	.0585651	-2.63	0.009	-.2691613	-.0392322
urban	.0566364	.0205007	2.76	0.006	.016393	.0968797
lrsel self	-.0252763	.0070231	-3.60	0.000	-.0390628	-.0114899
own	-.1064759	.0380061	-2.80	0.005	-.1810826	-.0318691
_cons	.8740558	.0816679	10.70	0.000	.7137399	1.034372

```
2 .
3 . predict yhat
   (option xb assumed; fitted values)
   (339 missing values generated)
```

What's wrong with the linear probability model?

Figure: Predicted Probability of Voting Labour from OLS Estimation



How to address limitations of OLS?

- Any continuous probability distribution defined over the real line would work.
- We use the normal because it's widely studied (which produces probit), and the logistic because it's mathematically convenient (logs—which produces logit).
- Ideologues might have reasons to prefer one to the other. If your results hinge on using one versus the other, you have problems.
- What we need is a probability model that looks like the following:

$$E[y|\mathbf{x}] = 0[1 - F(\mathbf{x}\beta)] + 1[F(\mathbf{x}\beta)] = \mathbf{F}(\mathbf{x}\beta)$$

The Logit Data Generating Process

The general problem with binary data is identifying a data generating process, a probability function, that maps our systematic component $E(y_i|X) = \mathbf{x}_i\beta$ into the unit interval, i.e., between 0 and 1.

$$Prob(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i\beta)$$

The logistic distribution is like a normal with longer tails (i.e., more extreme values are likely).

$$Prob(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i\beta}} = \Lambda(\mathbf{x}_i\beta)$$

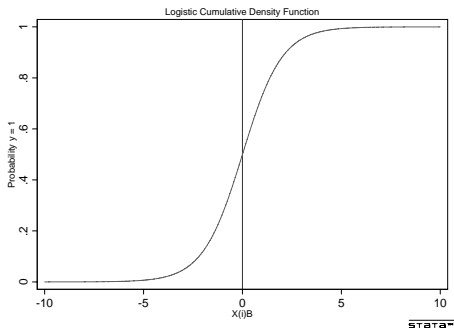
where Λ is the logistic cumulative distribution function.

The Logit Data Generating Process

We can generate example of logistic cumulative density function using Stata:

- `twoway function y=1/(1+exp(-x)), range(-10 10) xline(0) scheme(s1mono)`
- `ytitle("Probability y = 1", size(small))`
- `xtitle("X(i)B", size(small)) title("Logistic Cumulative Density Function", size(small))`

The Logit Data Generating Process



The Likelihood Function

- maximum likelihood provides a convenient and powerful method for estimating the parameters of the logit model.
- a key assumption is that the data are identically and independently distributed
- which allows us to form a likelihood function for the whole data from the product of the likelihoods for each observation:

$$(y_1, y_2, \dots, y_n) = P(y_1)P(y_2)\dots P(y_n)$$

$$= \prod_{y_i=1} F(\mathbf{x}_i\beta) \prod_{y_i=0} [1 - F(\mathbf{x}_i\beta)]$$

The Likelihood Function

In Likelihood notation:

$$L = \prod_{y_i=1}^N F(\mathbf{x}_i\beta)^{y_i} \prod_{y_i=0}^N [1 - F(\mathbf{x}_i\beta)]^{1-y_i}$$

- Each observation thus contributes something to the likelihood,
- either in the first part when $y_i = 1$,
- or in the second part when $y_i = 0$ (so $1 - y_i = 1$).

The Log Likelihood Function

As is typical with MLE, it is easier to work with the log-likelihood:

$$\ln L = \sum_{y_i=1}^N y_i \ln F(\mathbf{x}_i \beta) + \sum_{y_i=0}^N (1 - y_i) \ln [1 - F(\mathbf{x}_i \beta)]$$

- The only unknowns here are the vector of β
- But there is no simple analytic solution so this is typically accomplished iteratively.
- An example using the Labour incumbent voting data. Lets do a couple of iterations by hand.

The Likelihood Function

- In this example is Labour incumbent vote and takes on a value of 1 or 0.
- The independent variable, is income category that ranges in value from 10 (10,000) to 100 (100,000 or greater).

$$\ln L = \sum_{y_i=1}^{N=10} y_i \ln F(\alpha + \beta_1 \text{Income}) + \sum_{y_i=0}^{N=10} (1 - y_i) \ln [1 - F(\alpha + \beta_1 \text{Income})]$$

The Likelihood Function

- This can simply be calculated by hand
- Remember that the CDF for the logit is

$$F(\alpha + \beta_1 \text{Income}) = \frac{1}{1 + e^{-(\alpha + \beta_1 \text{Income})}}$$

which is equivalent to

$$F(\alpha + \beta_1 \text{Income}) = \frac{e^{(\alpha + \beta_1 \text{Income})}}{1 + e^{(\alpha + \beta_1 \text{Income})}}$$