

# Intermediate Social Statistics

## Lecture 3: Discrete choice

Raymond Duch

Nuffield College Oxford University

[www.raymond Duch.com](http://www.raymond Duch.com)

# 1 Motivation

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:

1. Presidential approval rates
2. Policy mood
3. Congressional polarization
4. Political tolerance
5. International trade
6. Globalization
7. Others?

But we have lots of dependent variables that we care a lot about that cannot be characterized as continuous, and those fall into several categories, such as (with examples):

1. Count (terrorist bombings)
2. Binary (votes)
3. Ordered (agree-to-disagree scales)
4. Multinomial (candidates in a primary; parties in multiparty election)

And we'll treat these separately.

All of our models will feel like probability models, of the sort:

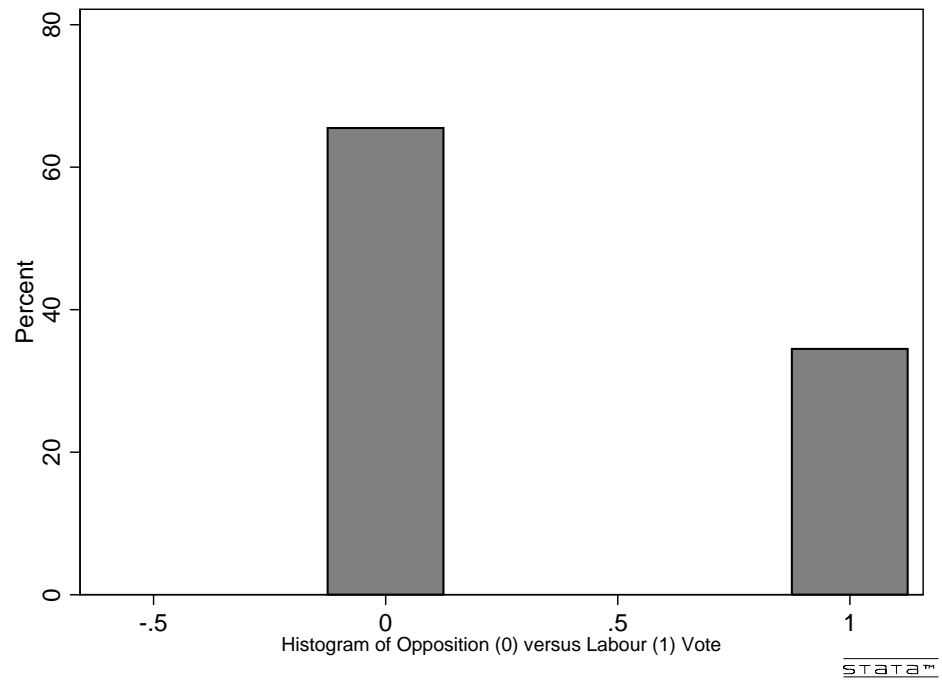
$$\text{Prob}(\text{event } j \text{ occurs}) = \text{Prob}(Y=j) = F[\text{stochastic component, systematic component}]$$

## 2 Binary choice models

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].

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.

Figure 1: Frequency of Labour versus Opposition Vote: UK 2004



### 3 What's wrong with the linear probability model?

Why not just estimate:

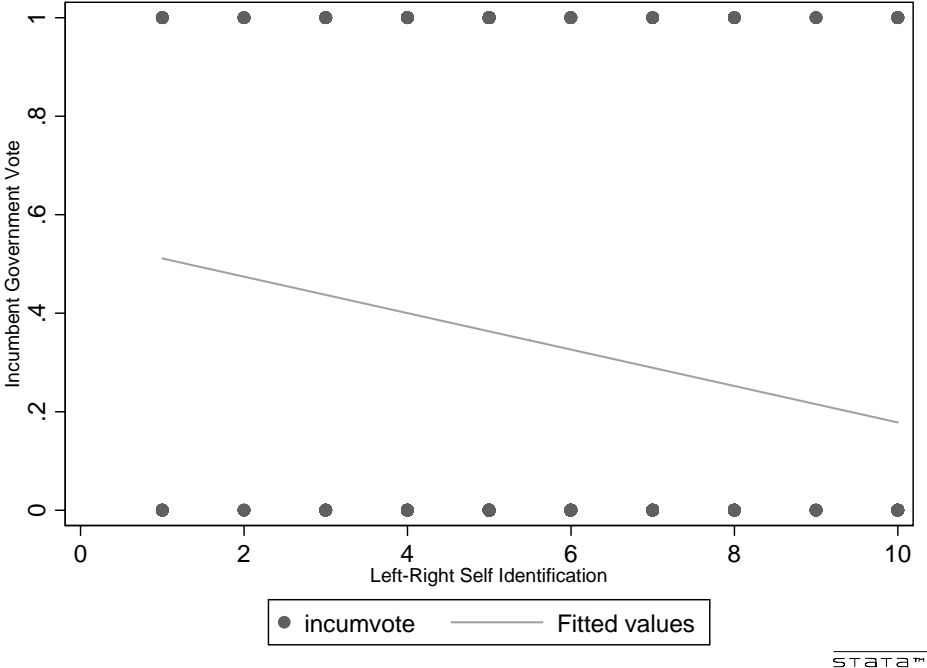
$$y = \mathbf{x}\beta + \varepsilon$$

where  $y = 0$  or  $y = 1$ ?

In terms of our example from the 2004 European Election study this would suggest,

$$\textit{LabourVote} = \textit{lrself} + \varepsilon$$

Figure 2: Vote Labour Incumbent versus Opposition by Class with OLS Regression Line

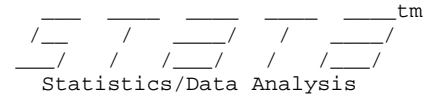


1. First, you can see where  $\varepsilon$  will be heteroskedastic. The variance of it will be lowest around  $p = 0.5$ , and highest close to 0 and 1. But we can fix this with GLS. So this isn't too too too serious.
2. Much more seriously, the model—you can see why—will make nonsense predictions, with  $p < 0$  and  $p > 1$ . That will also produce negative variances. We can see this more clearly by estimating the following model using OLS:

$$\textit{Labourvote} = \textit{retnat} + \textit{class} + \textit{union} + \textit{southwest} + \textit{urban} + \textit{lrsel} + \textit{own} + \varepsilon$$

Figure 3: Stata OLS Estimation of Labour Vote Model

Sunday January 27 06:45:29 2008 Page 1



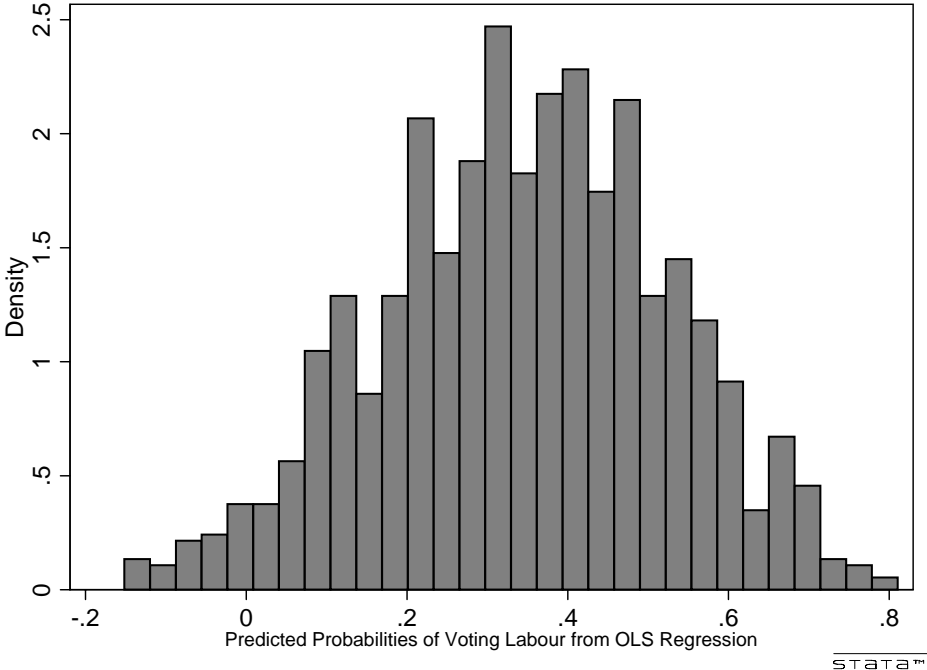
1 . regr incumvote retnat class union southwest urban lrself own

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

Figure 4: Predicted Probability of Voting Labour from OLS Estimation



What to do? 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:

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

## 4 Probability Models for Binary Data: Logit

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.

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

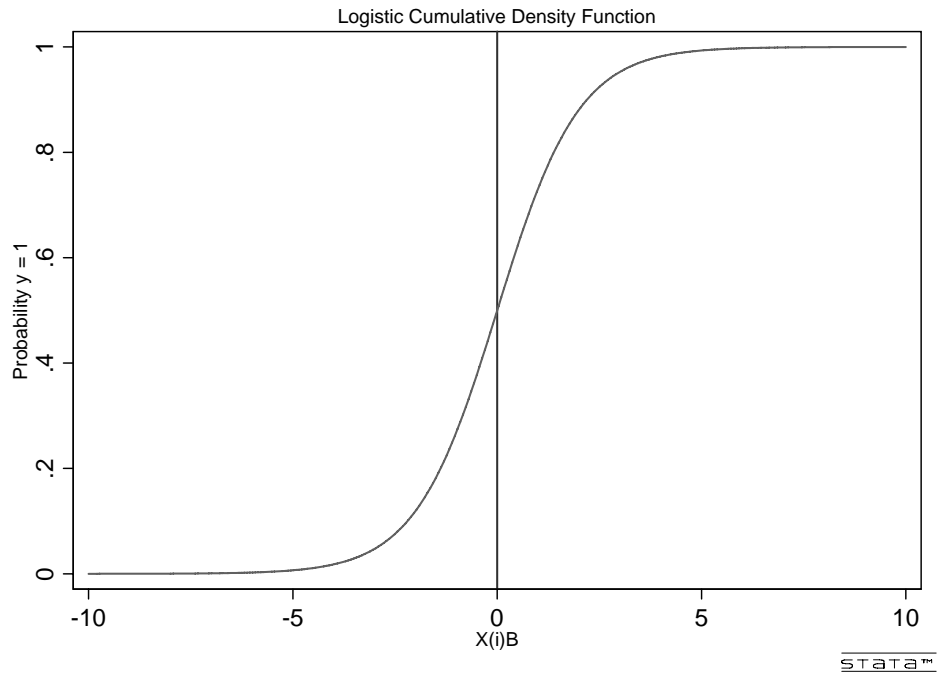
The logistic distribution is like a normal with longer tails (i.e., more extreme values are likely).

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

where  $\Lambda$  is the logistic cumulative distribution function.

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))
```



## 5 Estimation

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:

$$\begin{aligned}(y_1, y_2, \dots, y_n) &= P(y_1)P(y_2)\dots P(y_n) \\ &= \prod_{y_i=1} F(\mathbf{x}_i\beta) \prod_{y_i=0} [1 - F(\mathbf{x}_i\beta)]\end{aligned}$$

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  $1 - y_i = 1$ ). 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 is the vector of  $\beta$

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

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).

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

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

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

Lets try  $\alpha=-.02$  and  $\beta_1=1.2$

Table 1:  $\alpha=-.02$  and  $\beta_1=1.2$

Labour Vote	Income Category	Likelihood
1	10	-0.313
1	20	-0.371
1	30	-0.437
0	40	-0.913
0	50	-0.798
1	60	-0.693
0	70	-0.598
0	80	-0.513
0	90	-0.437
0	100	-0.371
Log Likelihood		-5.44

Lets try  $\alpha=-.05$  and  $\beta_1=1.0$

Table 2:  $\alpha=-.05$  and  $\beta_1=1.0$

Labour Vote	Income Category	Likelihood
1	10	-0.474
1	20	-0.693
1	30	-0.974
0	40	-0.313
0	50	-0.201
1	60	-2.126
0	70	-0.078
0	80	-0.0485
0	90	-0.0297
0	100	-0.0181
Log Likelihood		-4.958

Lets try  $\alpha=-.086$  and  $\beta_1=3.879$

Table 3:  $\alpha=-.086$  and  $\beta_1=3.879$

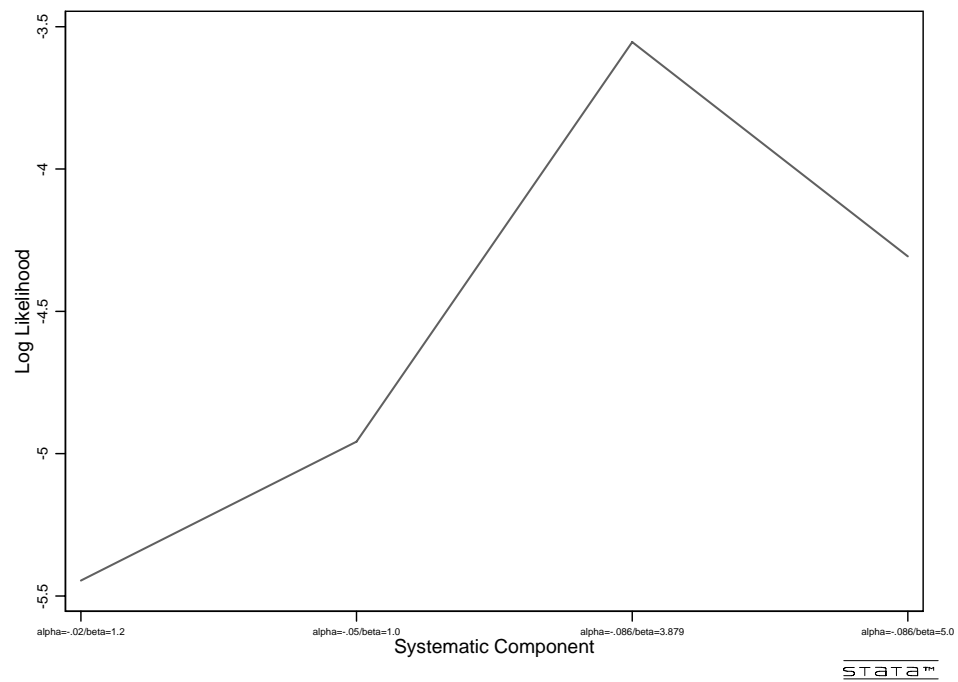
Labour Vote	Income Category	Likelihood
1	10	-0.047
1	20	-0.109
1	30	-0.241
0	40	-0.936
0	50	-0.505
1	60	-1.526
0	70	-0.111
0	80	-0.048
0	90	-0.021
0	100	-0.009
Log Likelihood		-3.55

Lets try  $\alpha=-.086$  and  $\beta_1=5$

Table 4:  $\alpha=-.086$  and  $\beta_1=5$

Labour Vote	Income Category	Likelihood
1	10	-0.016
1	20	-0.037
1	30	-0.085
0	40	-1.75
0	50	-1.103
1	60	-0.776
0	70	-0.308
0	80	-0.142
0	90	-0.063
0	100	-0.027
Log Likelihood		-4.308

Figure 6: Plot of likelihoods suggesting a maximum at  $\alpha=-.086$  and  $\beta_1=3.879$



## 6 Interpretation of Results

So if we're going to find the marginal effects of a change in some  $x$  on the probability of  $y = 1$ , we have to note that the parameters of the model,  $\beta$ , are not the marginal effect of  $x$  on  $y$ . In general:

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \left\{ \frac{dF(\mathbf{x}\beta)}{d(\mathbf{x}\beta)} \right\} \beta = f(\mathbf{x}\beta)\beta$$

And it's important to note that the values of the marginal effects (which is what the partials of  $y$  wrt  $x$  are) will change with  $x$ . That's what makes them nonlinear. (Think of a line. The marginal effect of  $x$  on  $y$  does not change anywhere over the course of that line. It is always  $\beta$ . Not here.)

Again, because these are nonlinear models, the marginal effects will be:

$$\frac{d\Lambda(\mathbf{x}\beta)}{d(\mathbf{x}\beta)} = \frac{e^{\mathbf{x}\beta}}{(1 + e^{\mathbf{x}\beta})^2} = \Lambda(\mathbf{x}\beta)[1 - \Lambda(\mathbf{x}\beta)]$$

or:

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \Lambda(\mathbf{x}\beta)[1 - \Lambda(\mathbf{x}\beta)]\beta$$

Interpretations of the parameter effects  $\beta$  are sensitive to the values of  $X$ .

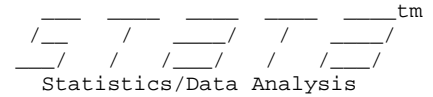
The most straightforward and typically most information is to generate some interesting predictions.

Here are the Stata logistic regression results for the UK 2004 model introduced at the beginning of the lecture

$$\textit{Labourvote} = \textit{retnat} + \textit{class} + \textit{union} + \textit{southwest} + \textit{urban} + \textit{lrself} + \textit{own} + \varepsilon$$

Figure 7: Stata Estimation of Logit Labour Vote Model

Tuesday January 29 13:11:38 2008 Page 1



```
1 . logit incumvote retnat class union southwest urban lrself own
```

```
Iteration 0: log likelihood = -507.77976
Iteration 1: log likelihood = -445.90699
Iteration 2: log likelihood = -443.71375
Iteration 3: log likelihood = -443.69615
Iteration 4: log likelihood = -443.69615
```

```
Logistic regression                                Number of obs =          785
                                                    LR chi2( 7)           =       128.17
                                                    Prob > chi2          =         0.0000
Log likelihood = -443.69615                       Pseudo R2            =         0.1262
```

incumvote	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
retnat	-.7038763	.1034497	-6.80	0.000	-.906634	-.5011186
class	-.3854525	.0861783	-4.47	0.000	-.5543589	-.216546
union	.5060247	.1915331	2.64	0.008	.1306267	.8814226
southwest	-.933915	.3541623	-2.64	0.008	-1.62806	-.2397697
urban	.3036703	.1060421	2.86	0.004	.0958315	.5115091
lrself	-.1339321	.0369293	-3.63	0.000	-.2063122	-.0615519
own	-.5231139	.1897608	-2.76	0.006	-.8950382	-.1511896
_cons	1.97143	.4262088	4.63	0.000	1.136076	2.806784

Using these estimates we can generate a predicted Labour vote probability for any respondent in our data set.

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

Here is a histogram of the predicted vote for all respondents in the UK 2004 survey based on each of their individual characteristics. How does this differ from the histogram of predicted votes generated by OLS regression?

Figure 8: Predicted Probabilities from Logistic Regression of Labour Vote

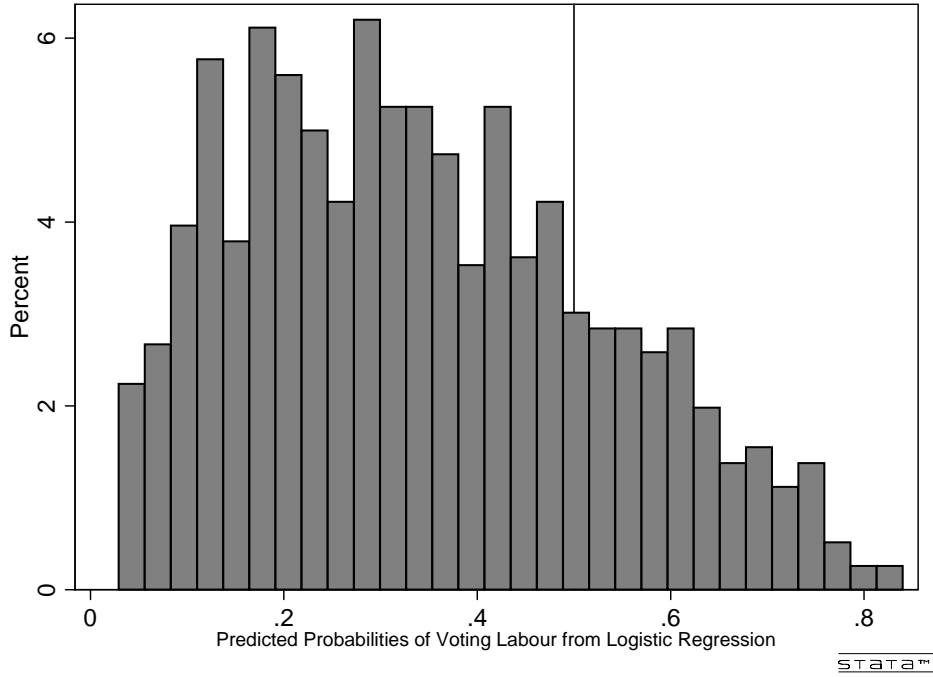
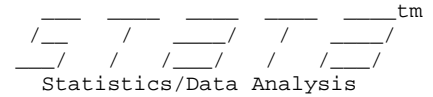


Figure 9: Variables in the Logit Model

Tuesday January 29 12:43:44 2008 Page 1



retnat | 1454 2.102476 .8110265 1 3

1 . tabulate retnat

retnat	Freq.	Percent	Cum.
better	411	28.27	28.27
same	483	33.22	61.49
worse	560	38.51	100.00
Total	1,454	100.00	

2 . tabulate class

class	Freq.	Percent	Cum.
working	600	43.20	43.20
lower middle	259	18.65	61.84
middle	461	33.19	95.03
upper middle	62	4.46	99.50
upper	7	0.50	100.00
Total	1,389	100.00	

3 . tabulate union

union	Freq.	Percent	Cum.
non-union	1,157	77.65	77.65
union	333	22.35	100.00
Total	1,490	100.00	

4 . tabulate southwest

southwest	Freq.	Percent	Cum.
0	1,370	91.33	91.33
1	130	8.67	100.00
Total	1,500	100.00	

5 . tabulate urban

urban	Freq.	Percent	Cum.
rural	493	33.11	33.11
small-medium town	576	38.68	71.79
large town	420	28.21	100.00
Total	1,489	100.00	

Generate meaningful predictions

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

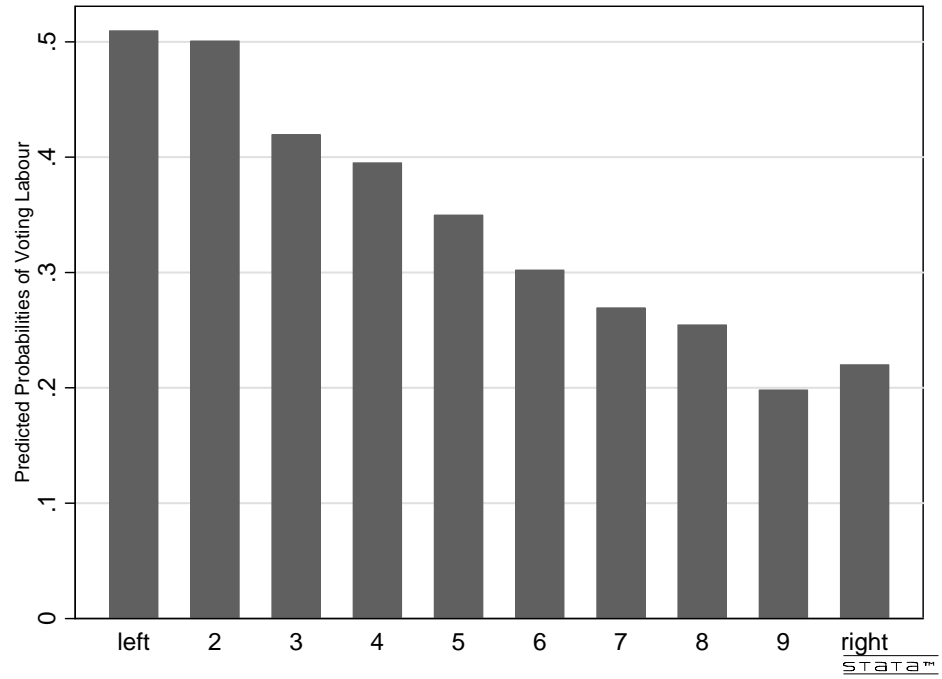
where we define a vector of  $x$  values for  $\mathbf{x}_i$  for a particular interesting case and use the  $\beta$  vector to generate the predicted probability of voting Labour for this particular "individual".

Table 5: Left-Wing Urban Working Class Male with Favorable View of Economy

Variable	Value	Coefficient	Value
Retnat	1	-.703	-.703
Class	1	-.385	-.385
Union	0	.506	0
Southwest	0	-.933	0
Urban	1	.303	.303
Lrself	1	-.133	-.133
Own	1	-.523	-.523
Constant	1	1.97	1.97
Systematic Component			0.529

$$\begin{aligned}
 Prob(y_i = 1|\mathbf{x}_i) &= \frac{e^{0.529}}{1 + e^{0.529}} \\
 &= .63
 \end{aligned}$$

Figure 10: Predicted Probabilities of Voting Labour by Left-Right Self Identification



Generating estimated changes in probabilities associated with meaningful changes in the independent variables. Here is an illustration of estimating the "economic vote" – the change in the probabilities of voting for the incumbent government (Labour) when economic evaluations shift one unit on a three-unit economic evaluation scale.

The economic vote for any individual ( $EV_i$ ) in the sample data set is the following:

$$EV_i = \frac{e^{\hat{\beta}_1(\text{Retnat}=1) + \sum_{j=1}^J \hat{\phi}_j Z_{ji}}}{1 + e^{\hat{\beta}_1(\text{Retnat}=1) + \sum_{j=1}^J \hat{\phi}_j Z_{ji}}} - \frac{e^{\hat{\beta}_1(\text{Retnat}=2) + \sum_{j=1}^J \hat{\phi}_j Z_{ji}}}{1 + e^{\hat{\beta}_1(\text{Retnat}=2) + \sum_{j=1}^J \hat{\phi}_j Z_{ji}}} \quad (1)$$

Table 6: Left-Wing Urban Working Class Male with Indifferent View of Economy

Variable	Value	Coefficient	Value
Retnat	2	-.703	-1.406
Class	1	-.385	-.385
Union	0	.506	0
Southwest	0	-.933	0
Urban	1	.303	.303
Lrself	1	-.133	-.133
Own	1	-.523	-.523
Constant	1	1.97	1.97
Systematic Component			-0.174

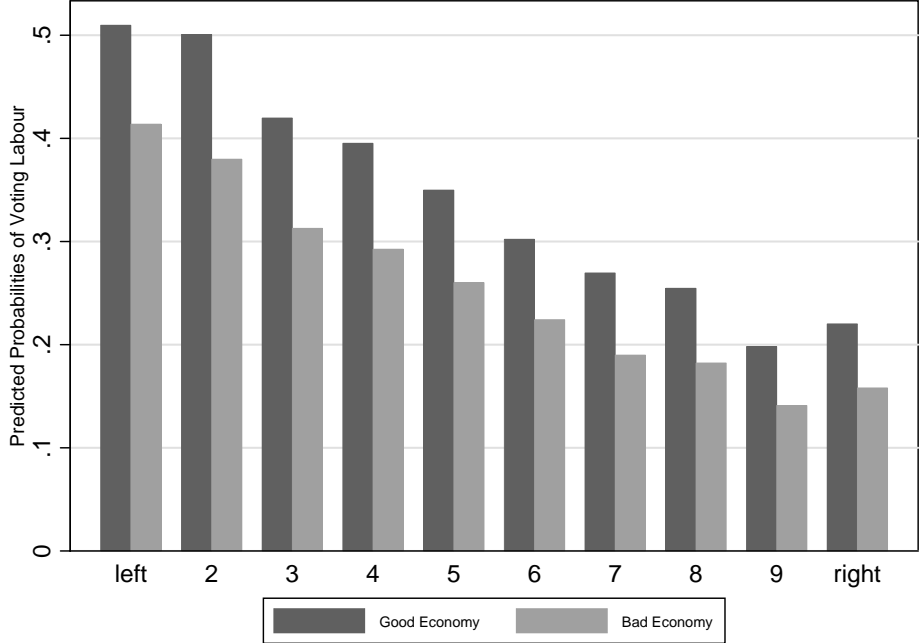
$$\begin{aligned} \text{Prob}(y_i = 1 | \mathbf{x}_i) &= \frac{e^{-0.174}}{1 + e^{-0.174}} \\ &= .46 \end{aligned}$$

$$\begin{aligned} EV_i &= \text{Prob}(y_i = 1 | \mathbf{x}_i) - \text{Prob}(y_i = 1 | \hat{\mathbf{x}}_i) \\ &= \frac{e^{0.529}}{1 + e^{0.529}} - \frac{e^{-0.174}}{1 + e^{-0.174}} \\ &= 0.63 - 0.46 \\ &= 0.17 \end{aligned}$$

Each of these estimated individual "economic votes" can be summarized for the whole sample by simply taking the average of these estimated changes in vote probabilities over the whole sample.

$$EV = \sum_{n=1}^N EV_i / N$$

Figure 11: Predicted Probabilities of voting Labour for Left-Right and Good-Bad Economic Perceptions



Stata™

## 7 Estimates of Uncertainty Associated with Probabilities/Changes in Probabilities

Estimating the uncertainty associated with these estimated probabilities is an essential part of interpreting our results. You of course are familiar with this in the case of political polling results. The confidence intervals on these estimates are typically very important in drawing conclusions regarding the lead of one candidate over another. For example we know that if in a two candidate race each candidate polls similar vote probabilities – say .49 versus .51 – that confidence intervals on these estimates lead us to conclude that there is no difference in the vote probabilities for the two candidates.

We also want to generate confidence intervals on our estimated changes in probabilities associated with manipulating the values of the independent variables in these equations. So from the example above, I estimated that a change in economic evaluations resulted in a 0.17 change in Labour vote probabilities (for this particular individual in the sample). Is this a big change – for example is it significantly different from a 0 change? We can estimate these standard errors using Clarify. Figure 12 presents an example.



## 8 Homework Questions

### Question 1

I have estimated a logistic model to explore how judicial selection procedures impact court rulings for or against the governor when he or she is a party in the case (governors are named in many cases). My main independent variable is whether the judge is elected or appointed. I also know whether the judge is the same party as the governor named in the case.

Data: 1000 rulings on cases in which governors are named as a party in the case. Drawn from 42 states over 10 years.

Dependent Variable: coded 1 for a ruling for the governor and 0 for a ruling against.

Independent variables: Elected = 1 if judge was elected, 0 if appointed Party = 1 if judge is same party as governor 0 if otherwise

Table 7: Results

Result	Coefficient
Elected	.07
Party	1 .61
Party*Elected	.21
Constant	.01

1. What is the stochastic component of this model?
2. What is the systematic component of this model?
3. What is the difference in the expected probability of ruling in the governors' favor between two judges who are both from the governor's party but one of whom is elected and the other appointed?
4. What is the difference in the expected probability of ruling in the governor's favor between two judges, both of whom are elected, but one who is from the governor's party and one who is not?
5. Comment on the strengths and weaknesses of this model in this situation

### Question 2

A five member committee votes 3-2 in favor of a proposal. Assume voting is independent. Let  $p$  be the probability that a committee member votes for the proposal.

1. We have no information with which to distinguish committee members (in the language of Bayesian statistics, we'd say that the committee members are exchangeable, but I digress). What is the maximum likelihood estimate (MLE) of  $p$ , the probability that any particular committee member votes for the proposal?
2. What is the log-likelihood of  $p = .5$ ? Compare this value of the log-likelihood function with that attained at the MLE with a likelihood ratio test. What does this say about the plausibility of  $H_0 : p = .5$ ?
3. How would your conclusion about the plausibility of  $H_0 : p = .5$  change if we observed
  - (a) i. a 10 person committee splitting 6-4 in favor of the proposal?
  - (b) ii. a 50 person assembly splitting 30-20 in favor of the proposal? i.e., what is happening to the likelihood function and/or the log-likelihood function in these cases relative to the case of a five person committee? In particular, what is happening to the 2nd derivative of the log-likelihood function in the neighborhood of the MLE?

### Question 3

Download the file `nagler.asc.dta` from my web site ([www.raymond Duch.com](http://www.raymond Duch.com)). This file contains 98,857 cases (welcome to large n research!) from the 1984 Current Population Survey, analyzed by Jonathan Nagler in two articles: *The Effects of Registration Laws and Education on Voter Turnout* *American Political Science Review*, 1991, 85:1393–1405; *Scobit: an alternative estimator to logit and probit* *American Journal of Political Science*, 1994, 38:230–255. The data in the file comprise the following variables (in column order): `turnout` 1 if the respondent reports turning out to vote in the 1984 presidential election, 0 otherwise. `educ` 1 for 0-4 yrs education; 2 for 5-7 yrs; 3 for 8 yrs; 4 for 9-11 yrs; 5 for 12 yrs; 6 for 1-3 yrs college; 7 for 4 yrs college; 8 for 5+ yrs college `age` age of respondent, in years `south` 1 if respondent lives in the South, 0 otherwise. `govelec` 1 if a gubernatorial election coincided with the presidential election closing number of days before election day that voter registration closes in the respondents state The following questions ask to you to estimate a series of logistic regression models. Construct a publication-quality table with the parameter estimates and standard errors for each the models, along with some summary information (e.g., goodness-of-fit, deviance, etc).

1. Estimate a logit model predicting turnout with the predictors `educ` and `age` and the square of each of these predictors. Provide a brief write-up of the parameter estimates (i.e., assess statistical significance and substantive implications) and the goodness-of-fit of the logistic regression model.
2. How many unique predicted probabilities are produced by this model? Explain how you derived your answer.

3. Compare the predicted probabilities from the logit model with the corresponding predicted probabilities from a probit model. How and why do they differ, if at all? Is there any statistical basis for preferring logit over probit or vice-versa?
4. Augment your logit model from the first part of this question with the following additional contextual predictors: south, govelec, and closing, and interactions between the two education variables (educ and educ2) and the closing date variable (i.e., make the effects of closing date quadratically conditional on the categorical education measure). Discuss the estimates and goodness-of-fit of this model in contrast with those obtained from the model for the previous question. Report a likelihood ratio test of the joint significance of the new predictors.
5. Using the estimates from the second model, plot the implied coefficient for closing as a function of education, given the interaction effects estimated above. Overlay 95 percent confidence intervals around the point estimates. Offer a substantive interpretation of what this plot reveals.
6. Using the estimates from the second model, consider a hypothetical nonsoutherner, in a state without a gubernatorial election, who has 12 years of education and has the median age of a non-southerner with 12 years of education. Plot the predicted probability of turnout for this person, as the closing date requirement varies over the range of closing date requirements observed in non-southern states. Overlay 95 percent confidence intervals around the point estimates.
7. Using the estimates from the second model, consider a hypothetical nonsoutherner, in a state without a gubernatorial election, who has 5+ years of college and has the median age of a non-southerner with 5+ years of college. Plot the predicted probability of turnout for this person, as the closing date requirement varies over the range of closing date requirements observed in non-southern states. Overlay 95 percent confidence intervals around the point estimates. Briefly compare the answers from this question with those from the previous question.

Due in class Wednesday, Week 5.