

**POL 502**  
**Lecture 9 Prof. Matthew Lebo**  
**April 7<sup>th</sup>, 2005**

Section 7.4 in the Text, Chapter 7

**Dummy Variables in Regression**

Dichotomous, qualitative variable

Reasons to think that the relationship we specify may be different based on something qualitative:

wartime versus peacetime

men vs. women

blacks vs whites

republicans vs. democrats.

We can create a dummy variable for any qualitative phenomenon.

1 when it occurs and zero otherwise.

Creating new variables in MINTAB is easy.

We can put these variables into a regression as we would any other type of variable.

Start looking at how a variable Y, depends on qualitative variables.

$$Y = \beta_D D_D + \beta_P D_P + \beta_L D_L + \varepsilon$$

Income is different for Doctors, Lawyers, and Professors.

We make a dummy variable for each.

The equation says an individual's income is given by the coefficient of his/her related dummy variable plus an error term.

So, for a lawyer the equation is....

for a doctor.....

for a professor.....

$\beta_D$  – average of all doctors' incomes.

A regression of just the above gives us this.

Is something missing from the equation?

Why is there no constant?

The intercept would equal the combination of the three variables – regression could not be run.

Usually, we include a constant and leave out one of the variables. Leave out lawyers (just like heaven).

$$Y = \beta_0 + \beta_D D_D + \beta_P D_P + \varepsilon$$

In this case for a lawyer the equation is simply:

$$Y = \beta_0 + \varepsilon$$

So the intercept is the average income of lawyers.

The variable we leave out is called our reference category.

We rethink our interpretation of the other coefficients:

They are now the differences between that category and the reference category.

ADDING another qualitative variable:

maybe gender has a role in our equation:

two ways to do this:

just add a new dummy variable for females:

$$Y = \beta_0 + \beta_D D_D + \beta_P D_P + \beta_F D_F + \varepsilon$$

Then we will have a value to add (or subtract) from expected income.

We leave out the male category = it plus female category would be a column of ones – perfect multicollinearity with constant would occur.

Second approach – get rid of the previous dummies and look at

Male doctor  
female doctors  
male professors  
female professors  
male lawyers  
female lawyers

Leave one out as a reference category:

$$Y = \beta_0 + \beta_{DF} D_{DF} + \beta_{DM} D_{DM} + \beta_{PF} D_{PF} + \beta_{PM} D_{PM} + \beta_{LF} D_{LF} + \varepsilon$$

Interpretation of the coefficients is easy – constant is for the reference category – male lawyers.

Each of the coefficients tells us the difference from Male lawyers.

What's the difference between the two ways of doing this?

The first says that the difference between men and women is the same regardless of profession. The second allows it to vary by profession.

The latter allows “INTERACTION EFFECTS”

Now, the equations we have looked at are unrealistic, all dummies.

Now move on to include another variable in the model.

More likely to have a quantitative variables in there also.

$$Y = \beta_0 + \beta_D D_D + \beta_P D_P + \beta_E E + \varepsilon$$

Where E is for experience.

Then we are taking account of experience and the dummies take account of the differences between each category from the reference category given experience.

Last equation is basically saying:

Income depends on experience, with a different intercept for each profession.

If we divided the data set into the three professions and regressed income on experience, we expect 3 parallel lines.

We don't do anything to account for differences in slope this way.

What if we think that the relationship between experience and income may be different for the 3 professions?

We can run the model separately for each.

OR,

We add special variables to account for the slope differences – INTERACTION TERMS:

$$Y = \beta_0 + \beta_D D_D + \beta_P D_P + \beta_E E + \beta_{ED}(D_D E) + \beta_{EP}(D_P E) + \varepsilon$$

Where  $D_P E$  simply means the  $D_P$  variable multiplied by the  $E$  variable.

So that  $\beta_{EP}$  tells us the difference in the slopes between Lawyers and Professors.

Now lines need not be parallel.

Substitute in zeros and 1s and see how this equation reduces for each of lawyers, doctors and professors.

**This is a great tool for non-linear data.**

When we suspect that the slope is different for different groups.

What are the reasons for including a dummy variable in a regression analysis?

When supported by theory.

What can they do?

- explain outliers perfectly
- minimizes error to ZERO
- increase model fit
- pull regression lines

Can be better to drop the case or set of cases.

Maybe better to split the data set.

If we had a dummy variable for every observation but one, we would have a perfect model – R-square of 1.

Practice Data set:

Year	Vote	GDP	PP	HS	Primary	VP	PPVP
1948	57.1	2.42	39	-54	1	1	39
1952	16.8	.07	32	-29	0	1	32
1956	86.1	.26	69	-18	1	1	69
1960	40.8	1.42	49	-47	1	0	0
1964	90.3	3.11	74	-5	0	1	74
1968	35.5	2.88	40	-48	0	1	40
1972	96.7	4.18	56	-12	1	1	56
1976	44.6	2.33	45	-48	0	1	45
1980	9.1	-1.38	21	-16	0	1	21
1984	97.6	3.95	52	-27	1	1	52
1988	79.2	1.91	52	-5	1	0	0
1992	31.4	1.4	40	-8	1	1	40
1996	70.4	3.6	51	-5	1	1	51
2000	46.7	3	59	3	1	0	0

Year is the year of the election.

V - share of the vote in the electoral college gained by the incumbent president (or his party's representative).

G - GDP growth that year.

PP - presidential popularity the month before the election.

HS - seats gained in the house by the (presidential) incumbent party (since the president's party always loses seats, this number is negative).

C - primary score. Scored 1 if the nominee wins at least 60 percent of the vote in the primaries and scored 0 if the nominee wins less than 60 percent.

1. What hypotheses would you test to explain V.

2. Professors Lewis-Beck and Rice believe that the Electoral College Vote Share of the incumbent can be explained using the following model:

$$V = \beta_0 + \beta_1 GDP + \beta_2 PP + \beta_3 HS + \beta_4 C$$

What are the hypotheses implied by this model? Be very specific - say what you are accounting for that is not part of each bivariate hypothesis.

Estimate the model.

Are the hypotheses supported by the model?

How much variance in vote share is explained by the model?

3. What percentage of the Electoral College vote would we expect an incumbent president to get if GDP growth was 2%, his party lost 30 seats in the House, his popularity level was 60%, and he had no problem winning the primaries (80% vote share)?

How about if GDP growth = -5%, PP= 30%, HS = +10, and he had a serious challenge in the primaries (55% vote share)?

This one's tricky:

According to the estimates of the model, holding all else equal, what level of GDP growth would have been necessary for George Bush to have won the 1992 election (i.e. have over 50% of the vote share in the Electoral College).

4. Collect data from reliable sources for 1996 and 2000. We will update the data set next class.

Explain each data point.....

1948 Dewey defeats Truman

1952 Eisenhower defeats Stevenson

1956 Eisenhower defeats Stevenson

1960 Kennedy defeats Nixon

1964 Johnson d. Goldwater

1968 Nixon defeat Humphrey

1972 Nixon d. McGovern

1976 Carter d. Ford

1980 Reagan d. Carter

1984 Reagan d. Mondale

1988 Bush d. Dukakis

1992 Clinton d. Bush

What other historical patterns can we see in the data?

President's party always loses house seats.

Recession of 1980 was huge.

Carter was extremely unpopular.

Post-Watergate fall in house seats.

Lyndon Johnson was very popular in 1964 and eventually was so unpopular that he didn't seek a second term.

- so popular, he barely lost seats in 1964.

3 biggest landslides occurred at time of 3 biggest economic booms.

What factors are important in deciding presidential elections?

Economy – gdp growth.

Presidential popularity

House seats – as a measure of overall partisan opinion in the US  
Whether or not a person is liked within their own party – success of campaign.

Dependent variable, share of the electoral college vote.  
- why is this good and bad?

$$V = \beta_0 + \beta_1 GDP + \beta_2 PP + \beta_3 HS + \beta_4 C$$

How can we reduce this equation for the case where Primary =1 and where Primary = 0.

What are the hypotheses implied by this model?

What signs should the coefficients be?

How good is this model for these data?

Answer question 3.

How can we see how well the model worked for each of the elections?

predict  $\hat{Y}$

see how well each year was predicted.

Why does the model not work so well for Gore?  
- what does the model not account for?  
- Any pattern to what happened for Gore?

**Over-predicts vote when the person running from the incumbent party is not the president.**

What does this tell us about the election and the model as a whole?

So, what's the problem of 2000?

Add a dummy variable for Vice president and an interaction for presidential approval \* vice president.