

POL 502
Lecture 10
Prof. Matthew Lebo
April 14, 2005

Last time we talked about dummy independent variables and interaction terms.

Review:

We can have dummy variables that will affect our constant, e.g.:

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

Or we can 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 β_{EP} tells us the difference in the slopes between Lawyers and Professors.

Now lines need not be parallel.

To create an interaction term we simply multiply one variable by another.

Neither one has to be a dummy variable.

We can multiply two interval level variables by each other.

This was tried in the first assignment by a couple of students.

Open the State Environment data set.

Does Waste always work as a variable?

Why should waste work? What is the process by which the amount of waste per person in a state should increase the degree of environmental voting?

High waste in a state with few people means that there may not be much pollution but there is a lot of industry.

So what really matters?

Waste in high density areas.

We can create an interaction variable that is Waste*Density.

Just go to the calculator function and create a new variable made up of Waste multiplied by Density.

Now we include that in a multivariate regression also with WASTE, INCOME, and ELECTION.

This gives us a very nice model and we can see that our interaction term is significant.

Creating the variable is no more difficult than when one was a dummy.

What is more difficult is the interpretation of the coefficient.

Its difficult to say what a one unit increase of waste*density means.

We can see that high values for either waste or density will give higher values for our interaction term which will lead to higher senate voting.

When waste is high but density is low people don't mind the pollution as much.

When waste is low but density high, there isn't much pollution.

The bottom line is that it is the interaction of the two variables that we want to capture.

Assignment 2

Option 2: Using the Country Health data set.

Look at the data set.

There are many extra variables that are only available for the wealthier countries in the data set.

A really complete analysis would use these variables for a study of the differences among developed countries as well as an additional model that measures the relationships among all the countries.

For example, what kind of relationship do we expect between the number of doctors and the infant mortality rate?

Is this true in the developed world?

Try regressing the infant mortality rate on docs, tcover, and stay.

This tells us more about the healthcare system – not just the number of doctors but the % of the total population covered by public health care and the length of hospital stays.

It might be a good idea to leave the U.S. out of your analysis because it has such outlying values.

Additional problems:

Some the variables have strange distributions.

There are so many poor countries and very big differences between the rich countries and the poor countries.

This can lead to non-linear relationships when we estimate models.

How can we solve this problem?

One possibility is to try to explain the non-linearity and to have our models account for it.

This can be done using dummy variables and interaction terms.

Another way is to use the logs of variables.

Logging a variable means instead of using the values of a variable we use the logarithms of the values.

This smushes the data points together and turns a wider spread into a more normal distribution.

It accentuates the differences between low values and minimizes the differences between the higher values.

What's the disadvantage?

We don't have a theoretical reason why the logs matter more than the variable itself.

We are forcing a non-linear relationship to be linear.

Its better than doing nothing but it would be best to try to model the non-linearity.

Presenting Statistical Results

We want to begin by outlining the phenomenon that we are explaining.

Explain how and when it occurs.

Explain how we have operationalized it in our data, how our sample was collected and then how it is distributed in our sample.

Histograms are useful.

Outline our model carefully.

What independent variables are important, how are they operationalized and distributed in the sample.

What are the hypothesized effects.

Graphs of the relationships or tables.

If there is an interaction going on, show it graphically.

E.g. a graph of life expectancy by gdp for developed countries versus non-developed.

Use proper table format for your output – what are the coefficients, standard errors and t statistics.

Include additional statistics and summaries of the regression.

Then write up your results.

Include a discussion of the substantive impact of each of your independent variables.

“We can be 95% certain that X has a significant impact on Y.”

“For each additional inch taller a person is, they are 5 pounds heavier, plus or minus 2 pounds.”

We get this from our confidence intervals.

Discuss any problems of sampling.

Discuss any problems of specification; multicollinearity, heteroskedasticity or serial correlation.

Discuss robustness of the model.

How does it react to the inclusion of additional variables or the exclusion of variables.

A relationship between one independent variable and the dependent variable is “robust” when you can change the model in many ways but still have the same relationship found.

This will be the case if there is a direct causal connection between them.

If there is not, including other independent variables that are more closely related will crowd out the independent variable.

For example, we may find that a variable of political opinion is well explained by ideology but that the affect of ideology changes a lot with the inclusion of other independent variables.

Party identification, on the other hand, is a more direct affect and is more robust to changes in the model.

Conclude with a discussion of the findings as well as possible weaknesses to your conclusions. Its best to anticipate criticism and answer it in advance.