Working with (and Testing) Dummies

I: Getting started (from an old exam)

Here are regression results from four models run using Wooldridge's ***ceosal1*** dataset,[1] predicting CEO salaries, as function of **roe** (average return on equity) and a bunch of industry dummies. There are four industry sector dummies in the dataset: **finance**, **utility**, **consprod**, and **indus**. They form an exhaustive set… so for each observation, one dummy variable is equal to 1 and the other three are 0.

```
. gen fin_roe = finance*roe
```

	(1) salary	(2) salary	(3) salary	(4) salary
finance	81.80	154.5	-14.14	654.1
	(0.36)	(0.67)	(-0.06)	(1.09)
roe		19.85	10.09	15.55
		(1.75)	(0.84)	(1.21)
utility			-601.6*	-555.4*
			(-2.18)	(-2.00)
fin_roe				-44.53
				(-1.21)
_cons	1263.1***	906.1***	1214.5***	1106.0***
	(11.73)	(3.94)	(4.53)	(3.91)
N	209	209	209	209
R-sq	0.001	0.015	0.038	0.045
adj. R-sq	-0.004	0.006	0.024	0.026

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

The first model regresses salary on a (0, 1) dummy for whether or not a firm is in the Finance sector. In the second model, ***roe*** has been added to the model. In the third model a (0,1) dummy variable for the Utility sector has been added in. And in the 4th model, an interaction between the finance dummy and roe has been added (so ***fin_roe*** = *finance*roe*).

1. Using the regression results and for each of the estimated models, write out expressions for the predicted salary for a CEO in the Utility industry, for a company having roe = 10. (No need to simplify; just write out the expressions.)

[1] Learn more about the dataset: http://fmwww.bc.edu/ec-p/data/wooldridge/ceosal1.des

Working with (and Testing) Dummies

II: Play Ball! *Winners and losers.*

The data used in this section are in a csv file, **teams.csv**, posted to the course website.[2]

Load the csv data into Stata, drop all observations prior to 1965 and generate **wprcnt** $\left(\%wins = \frac{w}{w+l}\right)$ and **rsra** $\left(\frac{runs\ scored}{runs\ allowed} = \frac{r}{ra}\right)$.

1. Regress **wprcnt** on **rsra**, and save your regression results with eststo**.**

Create a New York Yankees *intercept dummy* variable (**nyy**): **nyy = (franchID=="NYY")**

2. Add nyy to the OLS model you just estimated, save the results with eststo and capture the predicted values.

    a. On average, and controlling for rsra, would you say that the Yankees' %wins are higher or lower than for the rest of MLB. What is the estimated magnitude of that difference? Is it statistically significant?

    b. Generate a scatterplot of the predicted %wins for the Yankees and the rest of MLB (put rsra on the x-axis). Does the eyeball test say there's much difference?

3. Now, instead, interact the Yankees dummy with rsra (to create a *slope dummy*, **nyyrsra**) and rerun the regression model (with a slope dummy now, rather than an intercept dummy). As before, use eststo to save the results,

    a. Would you say that for the Yankees, increases in rsra are on average associated with higher or lower increases in %wins, compared to the rest of MLB? Is the difference in slopes statistically significant?

    b. And as above, generate a scatterplot of the predicted %wins for the Yankees and the rest of MLB. Does the eyeball test say there's much difference?

4. Finally, include both nyy and nyyrsra in your regression model. As before, save the results with eststo and capture the predicted values.

    a. Use the estimated equation to generate predicted %wins for the Yankees, when rsra=1. Write out an expression for the implied elasticity of changes in the predicted %wins for given changes in rsra, for the Yankees and at rsra=1. Just write out the expression, no need to simplify.

[2] Teams.csv was downloaded from Sean Lahman's website: http://www.seanlahman.com/baseball-archive/statistics/

b. Use an F-test to test the Null Hypothesis that the Yankees are no different from the rest of MLB, or more formally, that the true parameters for nyy and nyyrsra are both 0. Record your F statistic and *p value* (Prob > F) on the Answer Sheet. Do you accept or reject the null hypothesis? … at what level of statistical significance?

c. Use the regression results to write out an expression for the difference between the predicted %wins for NYY and for the rest of the league, as a function of rsra. Your expression should be a function of the estimated coefficients and rsra.

d. You are interested in how this difference changes as rsra increases, in differences-in differences (***diff-n-diff***). This will tell you the extent to which the Yanks are better able to convert increased rsra performance into increased %wins (scoring and preventing runs is one thing… but does it actually lead to more wins?).

   i. So differentiate your expression with respect to rsra, and record your answer on the Answer Sheet. Were you surprised by how easy this was?

   ii. Does your model predict that the difference in predicted %wins increases with rsra, or diminishes with rsra?

e. And as above, generate a scatterplot of the predicted %wins for the Yankees and the rest of MLB. Does the eyeball test say there's much difference?

5. Use esttab to summarize your regression results and attach that summary to your Answers.

Bring on the hapless NY Mets! (intercept effects)

Now, you'll be comparing the Yankees to their crosstown rivals, the New York Mets… allowing for separate intercept effects and testing for differences… in three ways.

Generate two new dummies: a team dummy for the NY Mets (**nym)** and a variable that is the sum of nym and nyy, so **nyny = (nyy+nym)**. You'll be testing the null hypothesis that the true parameter values for nyy and nym are the same.

6. There are several ways to do this (for each Way, record your Answers on the Answer Sheet):

   a. Way 1: **reg wprcnt nyy nym rsra**

      What is the predicted intercept for the Yanks? And for the Mets? Use the F test to test for differences between the Yanks and Mets effects: **test (nyy=nym)**. Do you reject the null hypothesis of no difference? … at what level of statistical significance?

b. Way 2: **reg wprcnt nyy nyny rsra**

What is the predicted intercept for the Yanks? And for the Mets? Use the t test (look at the t-stat and p-value for nyy) to test the null hypothesis that the true nyy parameter is zero. Do you reject the null hypothesis of no difference? … at what level of statistical significance?

Alternatively, use the F test to test for differences: **test nyy .** You should have found that the F stat was square of the t-stat and that the two associated probability values were the same. Did you?

c. Way 3: **reg wprcnt nym nyny rsra**

Repeat Way 2, now focusing on **nym**.

7. In each of the three cases you should have found the same F statistics and same F probabilities, as the tests are equivalent. Did you? What do you conclude?

III: Revisiting Nate's S&P Model

Looking at regional and Corruption Perceptions Index (CPI) effects.

You'll be working with the Nate Silver sovereign debt replication dataset, which you saw earlier in the semester. For convenience, I have posted that dataset in Excel format. You'll be exploring regional biases in the S&P ratings, as some have alleged that S&P is biased in favor of European and Asia/Pacific countries… or put differently, biased against the rest of the world.

1. Rerun your early replication of Nate Silver's *Sovereign Debt* model, now with the eurozone dummy added into the mix. Capture your results with eststo.

   reg NSRate lngdp corrupt inflation eurozone deficit_gdp debt_gdp

2. The estimated eurozone coefficient captures ***Eurozone*** effects in the model.[3] What does the estimated eurozone coefficient tell you about those effects? … compared to what? Do you find that S&P is biased in favor of Eurozone countries?

Now let's add additional regional intercept dummy variables to the model.

3. You'll see that I've added two regional intercept dummies (one for Asia/Pacific countries and the other for OtherEU countries) to the model. Those dummy variables were defined as follows (and yes, I realize that some of the calls are debatable) :

   Asia/Pacific (ap): Australia, China, Hong Kong, India, Indonesia, Japan, Korea, Malaysia, New Zealand, Philippines, Singapore, Taiwan, Thailand, and Vietnam.

[3] To brush up on your understanding of the Eurozone and the EU, I suggest: https://en.wikipedia.org/wiki/Eurozone

Working with (and Testing) Dummies

OtherEU (othereu): Bulgaria, Croatia, Czech Republic, Denmark, Hungary, Poland, Romania, Sweden and the United Kingdom.

Rerun your previous model, adding in ap and othereu, and capture the results with eststo.

a. Compare your coefficient estimates for the ***eurozone, ap*** and ***othereu*** dummies, and interpret the results (signs? magnitudes? statistical significance?)

b. How did the estimated eurozone coefficient change when you added these two additional dummies to the model? And why did that happen?

c. What happened to the parameter estimate for corrupt as we added these dummies?

d. Now use an F test to test for equality of eurozone and othereu effects… or put differently: Is it better to think of this as an ***EU*** effect rather than a ***Eurozone*** effect?

For the rest of this question, stay with this new model (with othereu and ap effects).

It turns out that the Corruption Perceptions Index (corrupt) is in fact an ordinal, not cardinal, variable… so only the relative magnitudes of the numbers have meaning… the particular values are otherwise meaningless. One way of handing ordinal RHS data is to create dummy variables.

As you saw in the "red meat" study, researchers sometimes employ quintile dummies to capture ordinal (and possibly nonlinear) effects. You can create quintile dummies for corrupt using the following command: **xtile corrupt5=corrupt, n(5)**

This will create a variable, ***corrupt5***, which takes on the values 1 through 5 reflecting the first quintile (value = 1; the bottom 20% of the values of corrupt), the second quintile (value = 2; the second 20% of the population distribution), and so forth. To see how this worked, plot the results: **scatter corrupt5 corrupt**

The choice to use quintiles here is arbitrary. There are more sophisticated methods for handling this issue of ordinal data… but not in this course.

4. To add quintile dummies for corrupt to your model, replace ***corrupt*** with ***i.corrupt5*** (the **i.** in front of corrupt5 will tell Stata to set up the quintile dummies), rerun the regression, capture the results with eststo, and plot the predicted values.

    a. Stata dropped one of the new dummies. Which quintile dummy was dropped? … and why?

    b. Interpret the results. What do the estimated parameter values for these new variables tell you?

    c. What happened to the parameter estimates for the eurozone, ap and othereu effects when you replaced corrupt with the quintile dummies?

d. In our previous analysis we assumed that there was a linear relationship between corrupt and NSRate. Looking at the scatter plot and the regression results, how do you feel about that assumption now?

e. Use esttab to print out a summary of the Part III results.