MULTIPLE REGRESSION IN EXCEL
EXCEL LAB #8

BUSN/ECON/FIN 130: Applied Statistics
Department of Economics and Business
Lake Forest College
Lake Forest, IL 60045
Copyright, 2013

Overview

This lab is written for Excel 2010, which is available to students in the library. The notation => can be read as "go to" or "click on." This notation will most often be used when navigating the menu or toolbars in Excel. To indicate a command or icon that you might click on or search for in Excel, **bold** will be used. Likewise, anything that you are to type into Excel will be **bolded** in the instructions. Do not enter such text as bolded text unless the instructions ask you to do so.

Tutorial

1. Open Excel, and install the Data Analysis Toolpak if necessary.

2. Open **House Values.xls**. Read through all 8 variables listed in the codebook to understand what each variable means and to find out how it is coded.

3. ***Creating Dummy Variables.***

   According to the codebook, there are two dummy variable (**garage** and **school**). According to the codebook (and by looking at your Excel file), however, you will notice that neither of the dummy variables are coded on a 0/1 scale. Regression analysis requires that dummy variables be recorded on a 0/1 scale, and it is best to do this according to the variable name itself. For example, the goal is to have **garage** equal 1 if the house has a garage and to equal 0 if it does not have a garage. The first task, therefore, is to make these necessary changes to **garage** and **school**.

   1. Insert a blank column between **garage** and **school**: **right click on column H (school)** => **Insert**.

   2. Enter **garage** in cell H1 (the new blank column) and enter **school** in cell J1 (the column to the right of **school**.

   3. To create the appropriate dummy variable for **garage**, notice that the codebook shows that **garage** equals 1 if the house has a garage, and equals 2 if the house does not have a garage. Thus, the "1" is correct in that is represents "yes," but we need to change the "2," which represents "no" to a "0". To do this, in H2 enter **=IF(G2=1,1,0)**.

Recall that the **IF** command requires a binary logic statement (in this case, does the value in cell G2 equal 1?) followed by the value to be assigned to a positive answer (in this case, 1) which is then followed by the value to be assigned to a negative answer (in this case, 0).

Copy and paste this formula down a few rows just to check that it is correct. You should notice that whenever column G equals 1, column H also equals 1; and whenever column G equals 2, column H equals 0. This shows that the formula is correct.

4. Copy and paste cell H2 in cells H3 through H5313.

5. Repeat steps 3 and 4 for the variable school. First check the codebook to see that the variable **school** in column I equals 1 if the house is located within 1 mile of a neighborhood school and equals 2 if not. Therefore, in cell J2 enter **=IF(I2=1,1,0)**. Copy cell I2 and paste it down through row 5,313.

As we don't need the original values, but the new values are linked to the original values, we need to copy and paste the new, formula-determined values, as their own values.

6. To paste the values from the formulas in column H into column H: **right click on column H** => **Copy** => **right click on column H again** => **Paste Special** => **Values** => **OK**. Now if you click on cell H2, you will see that the textbox includes the value of 0 rather than the formula. Repeat this for column I: **right click on column I** => **Copy** => **right click on column I again** => **Paste Special** => **Values** => **OK**.

7. Finally, as we don't need the original data values for the two dummy variables that are measured with 1's and 2's, delete these two columns (leaving one **garage** column with 0's and 1's and leaving one **school** column with 0's and 1's.

We are now going to estimate several multivariable regression models.

4. Create a new worksheet called **Model 1**, and place it to the right of **Houses Data**.

5. ***Regression Using the Data Analysis Toolpak.***

   In **Model 1**, we are going to regress housing value on square footage of the house (**sqfoot**), the number of rooms in the house (**rooms**), whether the house has a garage (**garage**), and whether the house is located near an elementary school (**school**).

   1. Copy the data from the **Houses Data** worksheet to the **Model 1** worksheet: in the **Houses Data** worksheet, **left click on column A and drag through column H** => **right click in the shaded area** => **Copy** => go to the **Model 1** worksheet => **right click on column A** => **Paste**.

   2. The first step in regression analysis in Excel is to delete all columns of data corresponding to variables that are not in the model (or, at least move them). As the data in **Model 1** include **value**, **sqfoot**, **rooms**, **garage**, and **school**, delete the columns of data associated with **lot**, **bedrms**, and **baths**: **left click on column B (lot)** => **press and hold the Ctrl key** => **left click on column E (bedrms)** => **left click on column F (baths)** => **right click anywhere in the shaded area** => **Delete**.

   3. To execute the regression: **Data** tab => **Analysis** box => **Data Analysis** => **Regression => OK** => the *y* variable is A1:A5313 => the *x* variables are B1:E5313 => click labels => Output Range = G1 => **OK**.

4. Reformat the coefficient estimates and standard errors (cells H17 through I21) to have no values after the decimal point as these estimates are for the dollar value of a house.

5. Reformat the t-stats and *p*-values (cells J17 through K21) to have 4 numbers after the decimal point as this is standard.

This part of the regression results should now look like the following.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-29474	8992	-3.2779	0.0011
sqfoot	39	3	14.4035	0.0000
rooms	10826	1359	7.9660	0.0000
garage	51670	5310	9.7314	0.0000
school	20263	3799	5.3344	0.0000

Notice that all of the variables in Model 1 are statistically significant. To interpret them:

1. Each additional square foot of living space in a house is expected to increase the value of the house by $39, all else equal.

2. Each additional room in a house is expected to be associated with a value that is $10,826 more than a comparable house without an additional room.

3. Having a garage is expected to raise the value of a house by $51,670.

4. Being located near an elementary school is expected to increase the value of a house by $20,263 compared to a comparable house that is not located near an elementary school.

6. Create a new worksheet called **Model 2**. Place it to the right of **Model 1**.

In **Model 2**, we are going to regress housing value on square footage of the house (**sqfoot**), the number of rooms in the house (**rooms**), the number of bedrooms (**bedrms**), whether the house has a garage (**garage**), and whether the house is located near an elementary school (**school**).

1. Copy the data from the **Houses Data** worksheet to the **Model 2** worksheet.

2. Delete the **lot** and **baths** variables as they are not in the regression model.

3. To execute the regression: **Data** tab => **Analysis** box => **Data Analysis** => **Regression => OK** => the *y* variable is A1:A5313 => the *x* variables are B1:F5313 => click labels => Output Range = H1 => **OK**.

4. Reformat the coefficient estimates and standard errors (cells I17 through J22) to have no values after the decimal point as these estimates are for the dollar value of a house.

5. Reformat the t-stats and *p*-values (cells K17 through L22) to have 4 numbers after the decimal point as this is standard.

This part of the regression results should now look like the following.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-32480	9422	-3.4474	0.0006
sqfoot	39	3	14.3134	0.0000
rooms	9733	1701	5.7212	0.0000
bedrms	3326	3114	1.0683	0.2854
garage	51384	5316	9.6652	0.0000
school	20058	3803	5.2737	0.0000

Notice that all of the variables in Model 2 are statistically significant except the number of bedrooms. Notice too that the estimated coefficients changed from Model 1 to Model 2. This is a very important lesson in multiple variable regression – the estimated coefficients depend on which variables are included in the model.

To interpret the estimated coefficients:

1. Each additional square foot of living space in a house is expected to increase the value of the house by $25, all else equal.

2. Each additional room in a house is expected to be associated with a value that is $9,733 more than a comparable house without an additional room.

3. Each additional bedroom in a house is expected to be associated with a value that is $3,326 more than a comparable house without an additional room.

4. Having a garage is expected to raise the value of a house by $51,384.

5. Being located near an elementary school is expected to increase the value of a house by $20,058 compared to a comparable house that is not located near an elementary school.

Before moving on, notice that **Model 2** includes all of the variables included in **Model 1** plus more. Consequently, it must be that the R-squared for **Model 2** is greater than that for **Model 1**, and it is as 0.1475 > 0.1473.

The last tutorial task is to investigate how coefficient estimates change when the variables are rescaled.

7. Create a new worksheet called **Model 3**. Place it to the right of **Model 2**.

8. In **Model 3**, we are going to repeat **Model 2** exactly, except that we will change the scale of some of the variables.

    1. Copy the data from the **Model 2** worksheet to the **Model 3** worksheet.

    2. The first change we want to make is to measure house values is thousands of dollars rather than in dollars. Right click on column A => **Insert**. The data should have all been moved one column to the right, and column A is now blank.

3. In cell A1 enter **value**. In cell A2 enter **=B2/1000**. Whereas the value of the house for the first observation is $25,000, cell A2 now has the value at 25. Copy and paste this for the entire column.

4. To remove the formula that is generating the values in column A: **right click on column A** => **COPY** => **right click on column A again** => **PASTE SPECIAL** => **VALUES**.

5. Delete column B (the original **value** column).

6. The only other change in variables that we are going to make is to the square footage of the house, **sqfoot**. Rather than measuring in square feet, we will measure square feet in 100s. Right click on column C (**sqfoot**) => **Insert**. The data should have all been moved one column to the right, and column C is now blank. In cell C1 enter **sqfoot**. In cell C2 enter **=D2/100**. Whereas the square footage of the house for the first observation is 2,000, cell C2 now has the value at 20. Copy and paste cell C2 down the entire column.

7. To remove the formula generating the values in column C, **right click on column C** => **COPY** => **right click on column C** => **PASTE SPECIAL** => **VALUES**. Column C now contains square footages measured in $1,000s.

8. Delete column D (the original **sqfoot** column).

9. Re-estimate **Model 2**, with these two measurement adjustments, as **Model 2**. In particular: **Data** tab => **Analysis** box => **Data Analysis** => **Regression => OK** => the *y* variable is A1:A5313 => the *x* variables are B1:F5313 => click labels => Output Range = H1 => **OK**.

10. Reformat the coefficient estimates, standard errors, t-stats, and *p*-values (cells I17 through L22) to all have 4 numbers after the decimal point.

This part of the regression results should now look like the following.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-32.4798	9.42	-3.4474	0.0006
sqfoot	3.9294	0.27	14.3134	0.0000
rooms	9.7332	1.70	5.7212	0.0000
bedrms	3.3265	3.11	1.0683	0.2854
garage	51.3836	5.32	9.6652	0.0000
school	20.0583	3.80	5.2737	0.0000

Notice that (except for **sqfoot**) all of the coefficient estimates for **Model 3** are exactly what they were for **Model 2**, with the exception that the decimal point has been moved over 3 places for both the coefficient estimate and the standard error. This happens because the data have not changed (i.e., the information contained in the variables included in the model has not changed), and so the coefficient estimates must predict the same thing as before. Whereas before, however, the coefficient estimates might have predicted a house

value to be $89,203, under **Model 3** the prediction will state the value as 89.203 (which is $89,203 measured in thousands). For example,

1. Under **Model 2**, an additional room is expected to increase house value by $9,733 above the value of a comparable house.

2. Under **Model 3**, an additional room is expected to increase house value by 9.733 thousands of dollars above the value of a comparable house.

The lesson is that when the dependent variable is scaled in some fashion, all coefficient estimates are scaled in an identical fashion. Notice too that the t-stats and the *p*-values for **Model 3** are exactly the same as for **Model 2**. This too must be the case as the information embedded in the data is the same between the two models. The R-squared for **Model 3** is also identical to the R-squared for **Model 2** at 0.1475 for each.

Lastly, notice that the estimated coefficient on **sqfoot** has changed from 39 to 3.9. Under Model 2, increasing the square footage by 1 square foot was expected to increase the value by $39. Thus, increasing the square footage by 100 square feet would be expected to increase the value by $3,900. As value is now measured in thousands, though, this 100 square foot increase is predicted to increase value by 3.9 thousands of dollars.

9. Save your work (4 worksheets in all) as YourName_Lab8_Tutorial.xlsx.

Exercises

1. Open **Pennsylvania Schools.xlsx**. A codebook for all 11 variables in **Pennsylvania Schools.xlsx** can be found at the end of the lab. Read the codebook carefully to learn what the variables are.

2. The only dummy variable in **Pennsylvania Schools.xls** is **MSA**, but it isn't measured on a 0/1 scale. Using the **IF** command, rescale all values so that **MSA** = 1 if the district is located in a Metropolitan Statistical Area and = 0 if the district is not located in a Metropolitan Statistical Area. Be sure to **copy and paste values** to avoid probablems of moving columns in which the **IF** formula was used.

3. Rescale **enroll** so it is measured in 100s. Again be sure to **copy and paste** to get rid of the potential formula problem.

4. Open a new worksheet. Call it **Model 1**. Place this worksheet to the right of **PA School Data**.

5. **Model 1**. For Model 1, regress district minimum salary (the *y* variable) on being in an MSA, enrollment, the rate of reduced-price or free lunch, and average SAT math score.
    1. Have the regression results begin in cell H2. Format the estimated coefficients, standard errors, t-stats, and *p*-values so they are easy to read.
    2. In cells H27 through H30, enter (right justified) **MSA**, **Enrollment**, **Lunch**, and **Sat Math** respectively.
    3. In cells I27 through I30 (center justified), state whether each estimated coefficient is a statistically significant predictor or not (enter "ssp" or "not ssp").
    4. In cells J27 through J30 (left justified), interpret each estimated coefficient.
    5. In cells H32 and H33 respectively enter **R-sq** and **Adj R-sq**. In cells I32 and I33, enter both statistics with 4 digits after the decimal place.

6. Open a new worksheet. Call it **Model 2**. Place this worksheet to the right of **Model 1**, and copy all of the data to it.

7. **Model 2**. For Model 2, regress district maximum salary (the *y*-variable) on being in an MSA, enrollment, the rate of reduced-price or free lunch, and average SAT math score.
    1. Have the regression results begin in cell H2. Format the estimated coefficients, standard errors, t-stats, and *p*-values so they are easy to read.
    2. In cells H27 through H30, enter (right justified) **MSA**, **Enrollment**, **Lunch**, and **Sat Math** respectively.
    3. In cells I27 through I30 (center justified), state whether each estimated coefficient is a statistically significant predictor or not (enter "ssp" or "not ssp").
    4. In cells J27 through J30 (left justified), interpret each estimated coefficient.
    5. In cells H32 and H33 respectively enter **R-sq** and **Adj R-sq**. In cells I32 and I33 both statistics with 4 digits after the decimal place.

8. Save your work (including all 6 worksheets) as YourName_Lab8_Exercises.xlsx.

Turning in your work

Email both files, YourName_Lab8.Tutorial.xlsx and YourName_Lab8_Exercises.xlsx, to your professor as file attachments to a single email with the subject heading Excel Lab 8: Your Name. Turn in your filled-in answer sheet during class.

Codebook for House Values.xlsx

Houses.xlsx contains 5,312 observations on individual houses. The data includes the following 8 characteristics on each house.

value	current market value of unit, actual value $20,000 to $999998
lot	square footage of lot, actual value 300 sq feet to 880,000 sq feet
sqfoot	square footage of interior of unit, actual value: 400 to 5200
rooms	number of rooms in unit, actual value: 3 to 19.
bedrms	number of bedrooms in unit, actual value 1 to 10
baths	number of full bathrooms in unit, actual value 1 to 7
garage	garage or carport included with unit; 1 Yes, 2 No
school	neighborhood elementary school within 1 mile; 1 Yes, 2 No

Codebook for Pennsylvania Schools.xls

Paschools.xlsx contains one observation for each of the 496 school districts in Pennsylvania from 1996. Included for each school district are the following 11 variables.

Name	District name
City	Name of the city in which the district office is located.
Salmin	The district's lowest paid teacher salary in 1996-97.
Salmax	The district's highest paid teacher salary in 1996-97.
Msa	Equals 20 if the district is not in a metropolitan statistical area (MSA); equals 30 if the district is in an MSA.
Enroll	The district's student enrollment for 1996-97.
Attrate	The district's 1996 average daily attendance rate, measured 0 to 100.
Lunch	The district's 1996 percent of students who receive free or reduced-price lunch, measured 0 to 100.
Dropout	The district's 1996-97 high school dropout rate, measured 0 to 100.
Satverb	The district's average SAT Verbal score in fall 1996.
Satmath	The district's average SAT Math score in fall 1996.

Answer Sheet for Lab #8: Multiple Regression in Excel

Name: _____

Provide your answers from the Exercise portion of the lab on PA school districts.

1. For Model 1, provide your answers to questions 5.3 and 5.4.

Variable	Sig/Not Sig	Interpretation
MSA		
Enroll		
Lunch		
SatMath		

2. For Model 2, provide your answers to questions 7.3 and 7.4.

Variable	Sig/Not Sig	Interpretation
Lunch		
SatMath		