Dale Berger, CGU
Source: http://wise.cgu.edu Guides and Downloads

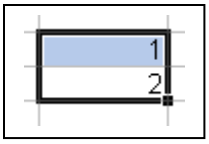
Regression Calculations with Excel

(Same data as SPSS Intro to Regression)

This is a demonstration of using Excel to do some basic regression calculations. The goal is to gain familiarity with both Excel and regression. Be sure that you are able to explain each step.

1. Open Excel on your computer. In the spreadsheet, enter **ID**, **X**, and **Y** in cells A1, B1, and C1, respectively. Enter the data as shown below.

	A	B	C	D
1	ID	X	Y	
2	1	3	8	
3	2	5	10	
4	3	6	4	
5	4	7	16	
6	5	9	12	
7				



Trick: When you have a sequence of numbers, as with ID, you can use a nifty feature in Excel to complete the sequence following the first two entries. After you enter 1 in A2 and 2 in A3, press Enter, and then move the cursor over the 1 in cell A2, hold the left mouse button and sweep down to highlight both cells A2 and A3. Release the mouse. You should see something like the figure at left, showing a bold rectangle around the two cells.

Note the little black square in the lower right corner of the bold rectangle. Move the cursor over this little black square – it will change shape into a + sign. Hold down the left mouse and sweep downward through cell A6. Release the mouse, and Voila! The time saved is more substantial when you have more numbers.

2. Prepare labels. In Cell A7 enter **Sum =**, in Cell A8 enter **Mean =**, and in Cell A9 enter **SD =.** You can right-justify these labels for better esthetics**.** Highlight the cells and click the Align Right icon, or right-click on the cell, Format Cells…, Alignment, under Horizontal select Right (Indent).

3. Compute the sum for X and for Y. Click View, Show, check Formula Bar. Move the cursor to Cell B7. Click ***fx*** to open the Insert Function window, select All as the category, select the function SUM, click OK, **highlight the X data in B2 through B6**, click OK. The sum is 30.

4. Now compute the mean and SD in a similar manner to Step 3. Move the cursor to Cell B8, select the function AVERAGE, etc. Move the cursor to Cell B9 and use the function STDEV.

5. We can compute the Sum, Mean, and SD for the Y variables using the formulas we created for X. Highlight the three cells B7, B8, and B9, move the cursor to the little black square on the bottom right of the box, hold the left mouse button and drag one column to the right. When we copy an Excel formula to a new cell, references to cells in the formula change accordingly. Check this by clicking on Cell B7 and then C7. You will see SUM(B2:B6) and SUM(C2:C6).

6. Now let's use Excel to examine the sources of variability in our data. If we did not have information on X, our best estimate for any Y score would be the mean of Y=10. In Cell D1 put the label **ybar**. In Cell D2 place the formula **=C$8**. Copy this formula to cells D3 through D6. The $ in front of the 8 'locks' the row reference to 8 so it doesn't change when **=C$8** is moved.

7. Now compute the error when we use the mean to predict Y. In Cell E1 enter **(y-ybar)**. In Cell E2 enter **=C2-D2**. Copy this formula to cells E3 through E6.

8. We can also calculate the squared errors. In Cell F1 enter **(y-ybar)^2**. In Cell F2 enter **=E2^2**. Copy this formula to cells F3 through F5.

9. Find the sums for the new columns. Copy the formula in Cell C7 to cells D7, E7, and F7.

10. An especially important concept is the Sum of Squares Total for Y. This is the sum of the squared errors in prediction, which is the sum of deviations from the mean shown in Cell F7 = 80 = SStotal. When you consider how it was calculated, you will recognize it as the numerator for the variance of Y. We can confirm this by entering into Cell C11 **=sqrt(F7/4).** Compare to Cell C9. Explain this result to Bumble. Hint: df = n-1; n=5

11. Now let's calculate the correlation between X and Y. In Cell A10 enter **r =**. Move the cursor to Cell B10. Click ***fx*** and select PEARSON, click OK. For Array1, highlight the X data in Cells B2 through B6. Move the cursor to Array2 and highlight the Y data in Cells C2 through C6. Click OK. The Pearson correlation is .45.

12. Excel has functions to compute the slope and the intercept. Put the labels **b =** into Cell A12 and **a =** into Cell A13. Put the cursor into Cell B12, click ***fx,*** select SLOPE, select the Y and X as in Step 6. Move the cursor to Cell B13, find the function INTERCEPT, select the Y and X data, etc.

13. Now we will explore some computations for regression. First let's find the predicted values for each Y. In Cell G1 enter **yhat**. In Cell G2 enter the prediction formula **= B$13 + B$12*B2**. Explain what this is to Bumble. Now drag this formula from Cell G2 into cells G3 through G6.

14. Next we can examine the errors in predicting Y when we use the regression model. First the labels: enter **(y-yhat)** into Cell H1, and **(y-yhat)^2** into Cell I1. Now the formulas: into Cell H2 enter **=C2-G2**, and into Cell I2 enter **=H2^2**. Compute the sums for these three columns by dragging the summation formula from Cell F7 into cells G7 through I7.

15. Notice that the sum of the squared errors from the regression model (SSerror) shown in Cell I7= 63.2 is somewhat smaller than the SStotal = 80 shown in Cell F7. Most, but not every regression prediction is closer to Y than the mean – see Case 2 where Y=10. However, on average, the squared errors are smaller from the regression model than from the mean. SSerror corresponds to the part of SStotal that **cannot** be predicted from the regression model.

16. If the regression model is useful, then the predicted values differ from the mean. We can compute errors for these values with Excel. First the labels: In Cell J1 enter **(yhat-ybar)** and in Cell K1 enter **(yhat-ybar)^2**. In Cell J2 enter **=G2-D2** and in Cell K2 enter **=J2^2**. Copy the formulas from J2 and K2 down the columns into cells J3 to J6 and K3 to K6. Finally, copy the formula for sums into cells J7 and K7.

17. The sum of squared deviations of the predicted values from the means shown in Cell K7 is the Sum of Squares Regression = SSreg. This corresponds to the portion of the SStotal that can be predicted by the regression model. SSreg/SStot = 16.2/80.0 = .2025 = r^2 = $.45^2$ = .2025.

Summary

	A	B	C	D	E	F	G	H	I	J	K
1	ID	X	Y	ybar	(y-ybar)	(y-ybar)^2	yhat	(y-yhat)	(y-yhat)^2	(yhat-ybar)	(yhat-ybar)^2
2	1	3	8	10	-2	4	7.3	0.7	0.4900	-2.7	7.29
3	2	5	10	10	0	0	9.1	0.9	0.8100	-0.9	0.81
4	3	6	4	10	-6	36	10	-6	36.0000	0	0.00
5	4	7	16	10	6	36	10.9	5.1	26.0100	0.9	0.81
6	5	9	12	10	2	4	12.7	-0.7	0.4900	2.7	7.29
7	Sum =	30	50	50	0	80	50.00	0	63.8000	0	16.2000
8	Mean =	6	10								
9	SD =	2.236	4.472								
10	r =	0.4500									
11			4.47214								
12	b =	0.9									
13	a =	4.6									

If we had no information on X and we were forced to guess each Y value, our best guess would be the mean of Y = 10 for each case as shown in column D. The amount of error in this prediction is shown in column E. The sum of the squared deviations from the mean in an index of total error in our predictions. This value is shown in Cell F7 as 80.0. This term is called the Sum of Squares Total = SStot. Note that it is the numerator for calculating the variance of Y.

If we know the value of X for each case, we may be able to use the regression model to make better predictions as shown in Column G. The errors in these predictions are shown in Column H, and the sum of the squared errors in prediction is shown in Cell I7 to be 63.8. This term is called the Sum of Squares Error = SSerr, and it corresponds to the amount of the total sum of squares that **cannot** be predicted with the regression model.

The amount of the total sum of squares that **can** be predicted from the model is found in Column K, where the sum is shown in Cell K7 as 16.2. This term is called the Sum of Squares Regression = SSreg.

$$SS_{tot} = SS_{reg} + SS_{err} ; \quad 80.0 = 16.2 + 63.8.$$

The proportion of the Sum of Squares Total that can be predicted from the regression model is (Sum of Squares Regression) / (Sum of Squares Total) = 16.2 / 80.0 = .2025. This is equal to the square of the Pearson correlation ($.45^2 = .2025$).

This relationship is the basis for the statement that r squared is the proportion of variance in Y than can be explained by X.

It is good practice to document files well, including your name and date on everything you do.

DB 160724