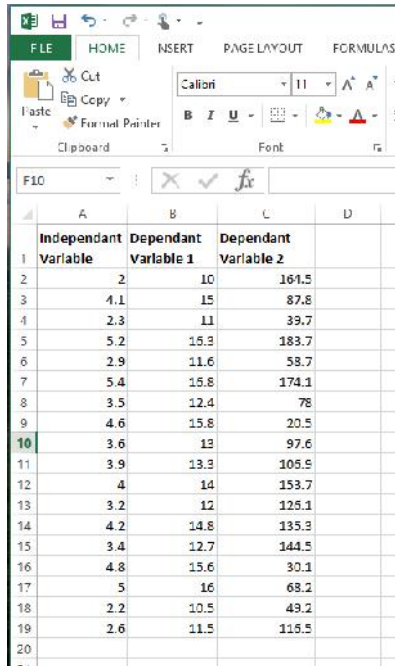


## OLS Regression and ANOVA in Excel 2013

**Note:** Other versions of excel generally have these same functions, but the menus/paths may differ

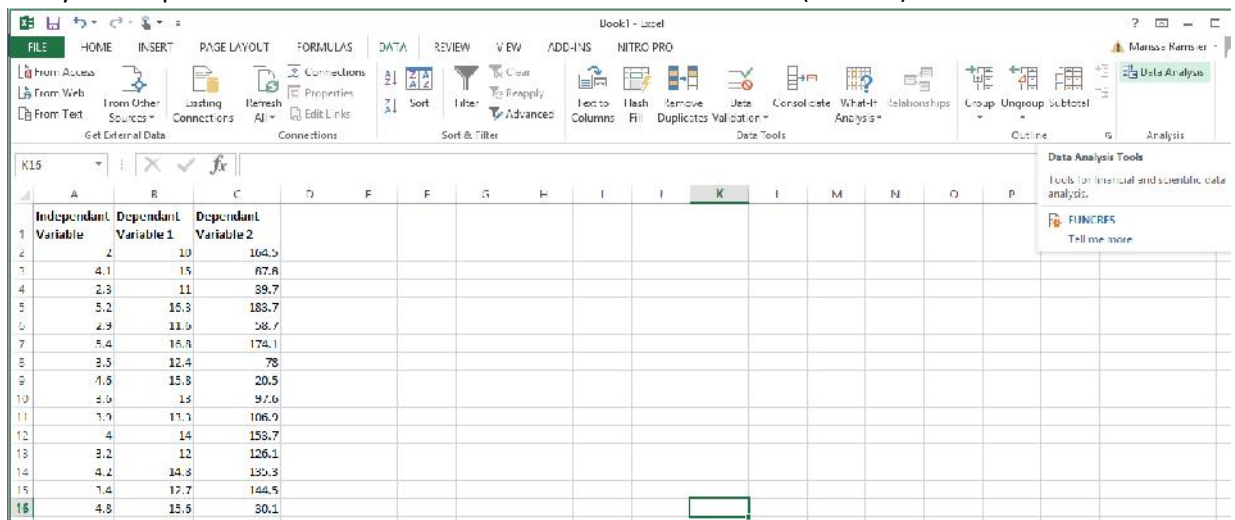
**Note:** Regression and ANOVA is appropriate if you are working with two continuous variables (e.g., 1.1, 1.2, 1.4, 2.7). Discrete numerical categories (e.g., 1, 2, 3, 4) can sometimes be treated as continuous. ANOVA is also appropriate if you have a categorical independent (e.g., yes/no coded as 1, 2) and a continuous dependent variable. Neither is appropriate if you the independent variable is continuous.

- (1) **Input data into excel.** Make sure data are formatted as numbers with no additional information in the cells:



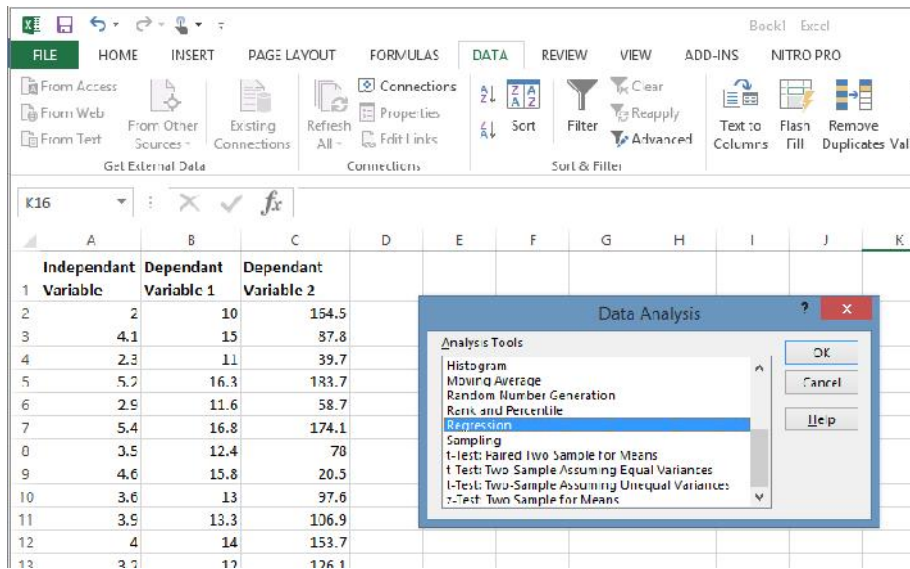
	A	B	C	D
1	Independent Variable	Dependent Variable 1	Dependent Variable 2	
2	2	10	161.5	
3	4.1	15	87.8	
4	2.3	11	39.7	
5	5.2	15.3	183.7	
6	2.9	11.6	58.7	
7	5.4	15.8	174.1	
8	3.5	12.4	78	
9	4.6	15.8	20.5	
10	3.6	13	97.6	
11	3.9	13.3	105.5	
12	4	14	153.7	
13	3.2	12	125.1	
14	4.2	14.8	135.3	
15	3.4	12.7	144.5	
16	4.8	15.6	30.1	
17	5	16	68.2	
18	2.2	10.5	49.2	
19	2.6	11.5	115.5	
20				

- (2) **Make sure you have the Data Analysis Toolpak.** Go to the DATA tab, and select the Data Analysis Toolpak (on the right). If you do not see that option, do a google search for “Data Analysis Toolpak in Excel” to find the instructions on how to add it (for free)

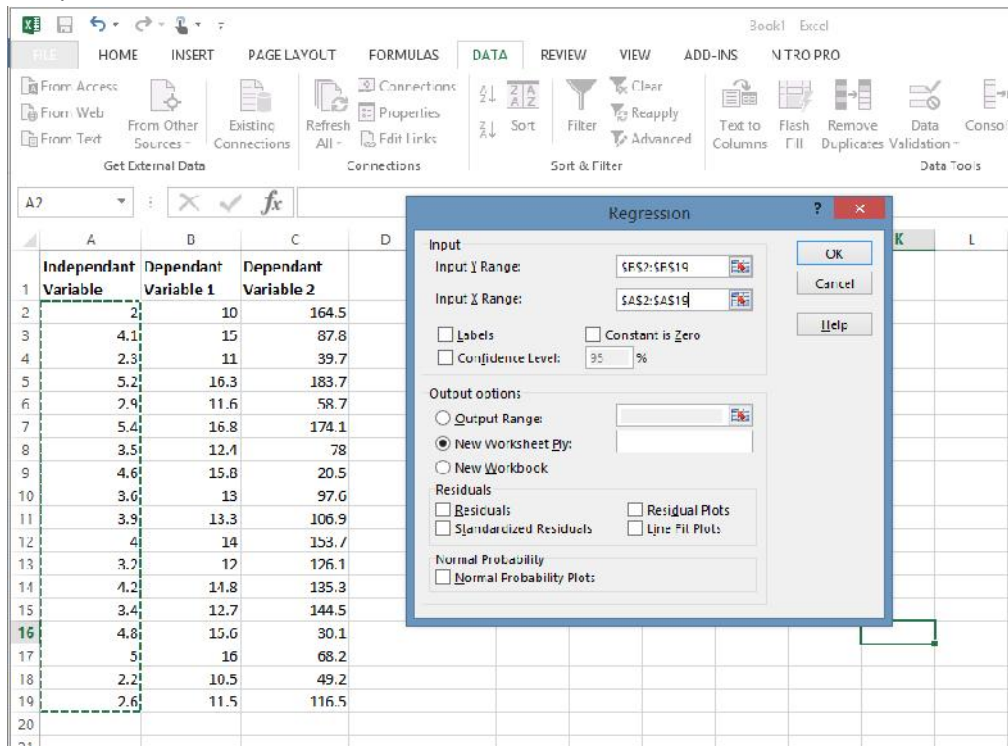


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Independent Variable	Dependent Variable 1	Dependent Variable 2													
2	2	10	164.5													
3	4.1	15	87.8													
4	2.3	11	39.7													
5	5.2	15.3	183.7													
6	2.9	11.6	58.7													
7	5.4	15.8	174.1													
8	3.5	12.4	78													
9	4.6	15.8	20.5													
10	3.6	13	97.6													
11	3.9	13.3	106.9													
12	4	14	153.7													
13	3.2	12	126.1													
14	4.2	14.8	135.3													
15	3.4	12.7	144.5													
16	4.8	15.6	30.1													
17																

(3) In the Data Analysis popup window, scroll down to select “Regression” and then “OK”



(4) Put your mouse in the box for “Input Y Range” and then drag a box around one of your dependent variables. Then put your mouse in “Input X Range” and drag a box around your independent variable. Then click “OK”



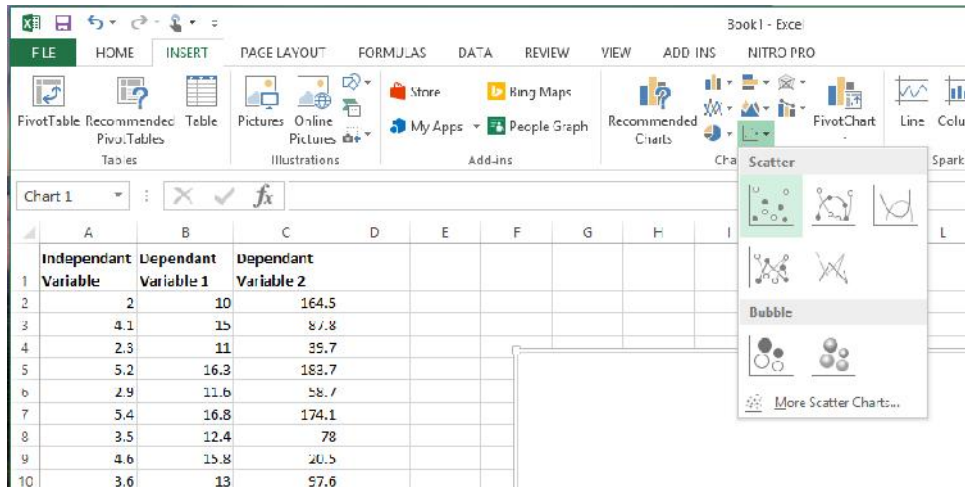
- (5) You should see the output of the regression analysis pop up in a new worksheet. It will look like the output below. Expand the width of column A if needed so you can see the text. (note: I added the blue highlighting)

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.985355735								
R Square	0.971004755								
Adjusted R Square	0.969192552								
Standard Error	0.376793048								
Observations	18								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	1	75.07120976	76.07121	535.8146	9.93995E-14				
Residual	16	2.271569017	0.141973						
Total	17	78.34277878							
		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept		5.96985774	0.33592351	17.7891	5.78E-12	5.25646177	6.581309777	5.25846177	6.661309777
X Variable 1		2.015576324	0.087074596	23.14767	9.94E-14	1.830566214	2.200166434	1.830966214	2.200166434

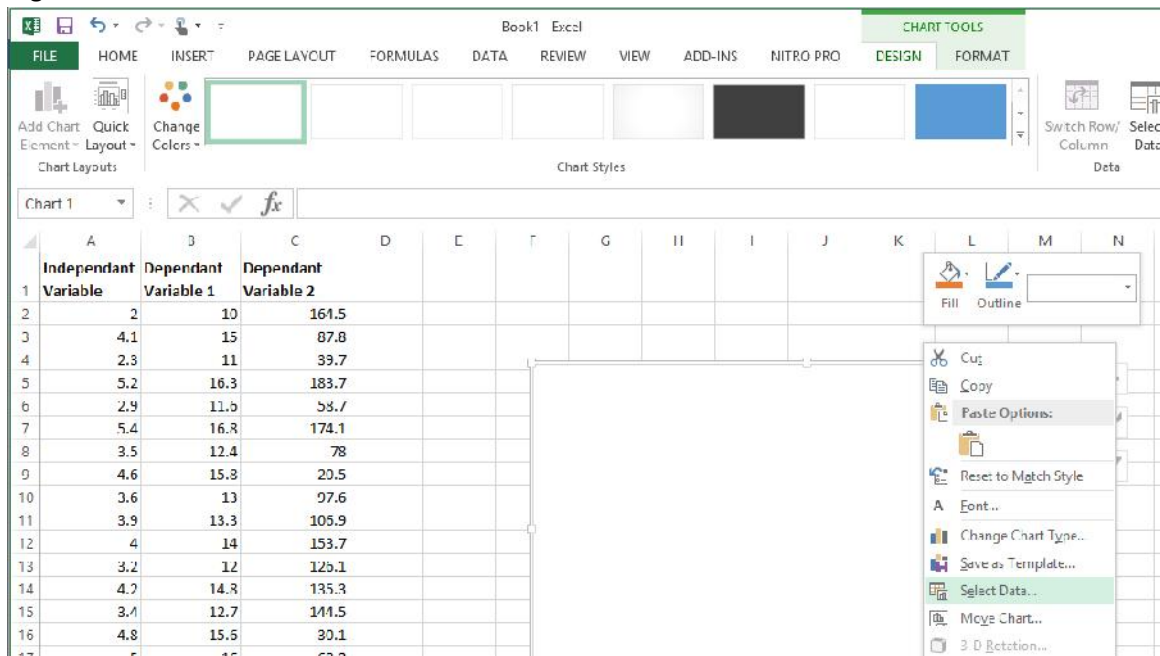
- (6) We are particularly interested in five items on the output (highlighted above):
- The R Square. This is how linear the relationship between the variables are. The closer to 1, the more linear the relationship, and the more variation in Y (dependent variable) that is explained by X (independent variable). In the example below, the R Square is 0.97 (round to 2 decimal places). This means the relationship is highly linear (“a good fit”), and 97% of variation in X is explained by Y. Remember that this does NOT reflect causation (we can’t know from this analysis if one variable causes variation in the other), but instead just shows there is a relationship.
  - Excel runs an “Analysis of Variance” (ANOVA) along with the OLS regression.
    - The F is a measure of the strength of the relationship. In the example, the F is 535.81, which is good. Don’t worry too much about the F, just write it down.
    - The Significance F is the P-Value. This tells us how significant the relationship is. In the example, the P-Value is very low, 9.39 E-14, which is excel’s way of using scientific notation. This means  $9.39 \times 10^{-14}$ , or 0.0000000000000939. It is excessive to report all this though. The convention is that the relationship is “statistically significant” if  $P < 0.05$ , and “highly statistically significant” if  $P < 0.01$ . So here you can simply record  $P < 0.01$ . If you want to show it is very very good, just report one more zero ( $P < 0.001$ ).
  - The coefficients give you the formula of your regression line. The formula of a line is:  $y = mx + b$ , where m is the slope and b is the y-intercept. In the excel output, the slope is the X Variable 1 coefficient, and the y-intercept is the Intercept Coefficient. So for the example,  $y = 2.0156x + 5.9699$ . I have rounded to 4 places, but 2 is usually OK also. This is our predictor equation. So, if we wanted to predict variable Y with a value of X, we would simply insert the X value into the equation and get Y.

(7) This is how you would state your results in a paper. There was a statistically significant relationship between (insert name of one variable) and (insert name of the other variable) ( $R^2=0.97$ , ANOVA  $F=535.81$ ,  $P<0.001$ ).

(8) Now, let's represent the data graphically. Go back to your data, and click the INSERT tab up top, then charts>scatterchart



(9) Right click on the chart area and "Select Data"



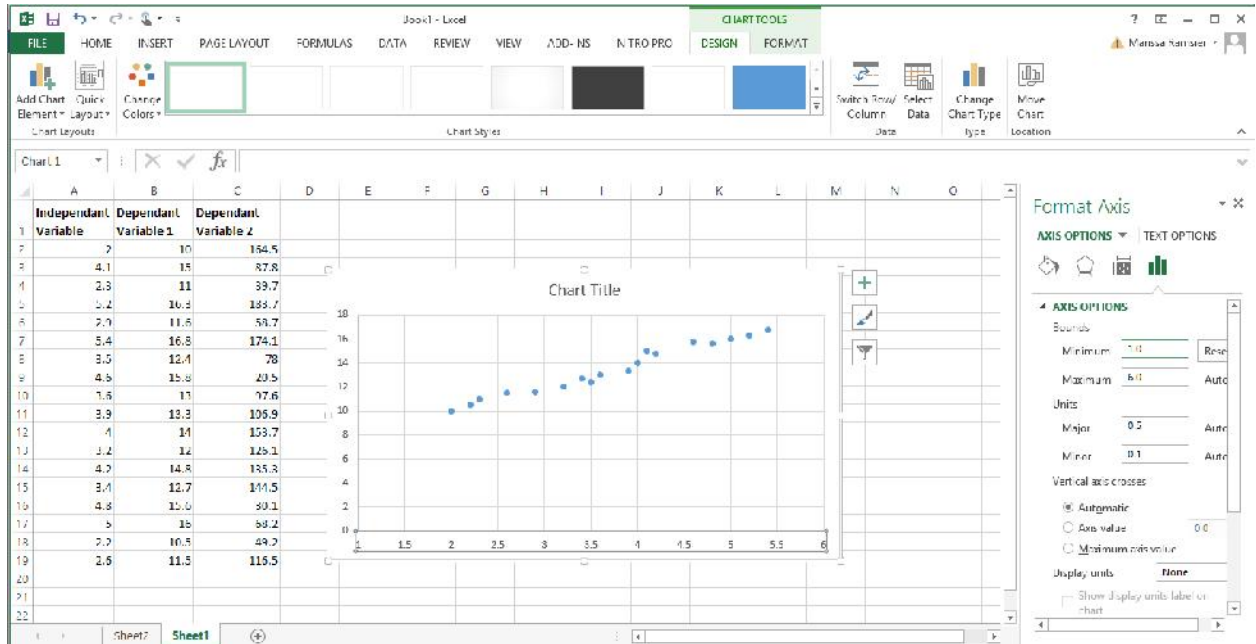
(10) Put your mouse in “Chart Data Range” and then draw a box around both columns of data (they will need to be next to each other).

	Independent Variable	Dependent Variable 1	Dependent Variable 2
2	2	10	154.5
3	4.1	15	87.8
4	2.3	11	39.7
5	5.2	16.3	183.7
6	2.9	11.6	58.7
7	5.4	16.8	174.1
8	3.5	12.4	78
9	4.5	15.8	20.5
10	3.5	13	97.6
11	3.9	13.3	106.9
12	4	14	153.7
13	3.2	12	126.1
14	4.7	14.8	133.3
15	3.4	12.7	144.5
16	4.8	15.6	30.1
17	5	16	68.2
18	2.2	10.5	49.2
19	2.5	11.5	116.5

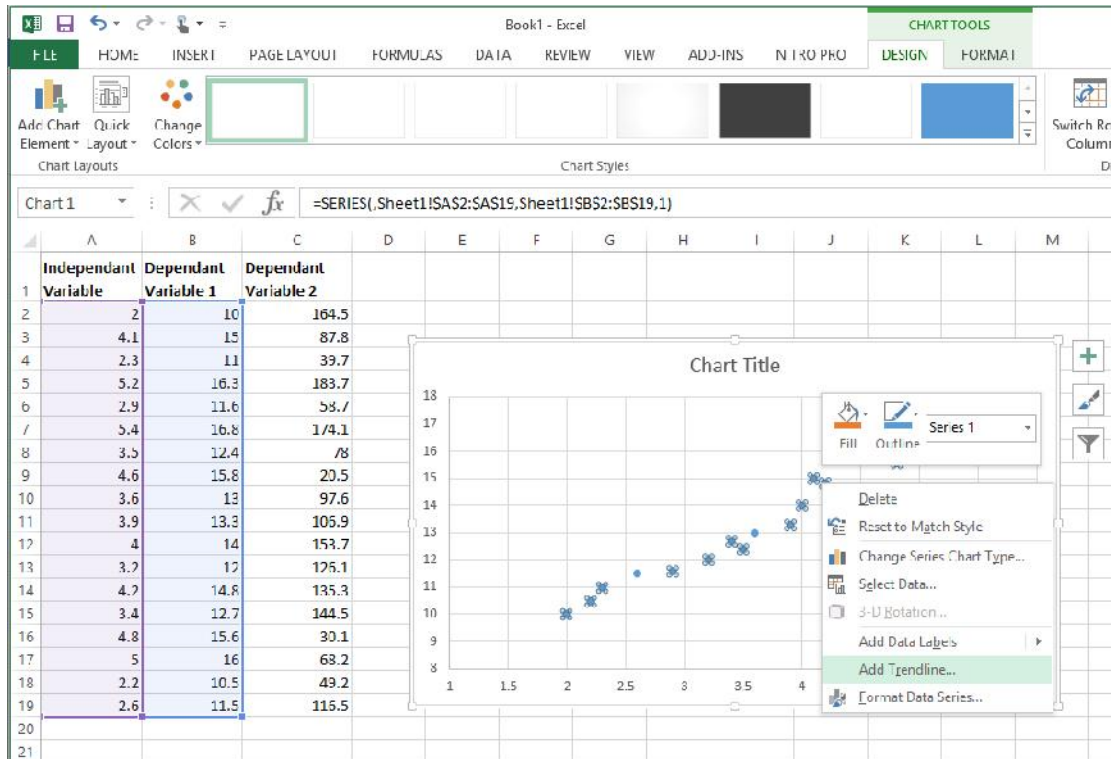
(11) Your data should now be in the chart:

	Independent Variable	Dependent Variable 1	Dependent Variable 2
2	2	10	154.5
3	4.1	15	87.8
4	2.3	11	39.7
5	5.2	16.3	183.7
6	2.9	11.6	58.7
7	5.4	16.8	174.1
8	3.5	12.4	78
9	4.5	15.8	20.5
10	3.5	13	97.6
11	3.9	13.3	106.9
12	4	14	153.7
13	3.2	12	126.1
14	4.7	14.8	133.3
15	3.4	12.7	144.5
16	4.8	15.6	30.1
17	5	16	68.2
18	2.2	10.5	49.2
19	2.5	11.5	116.5

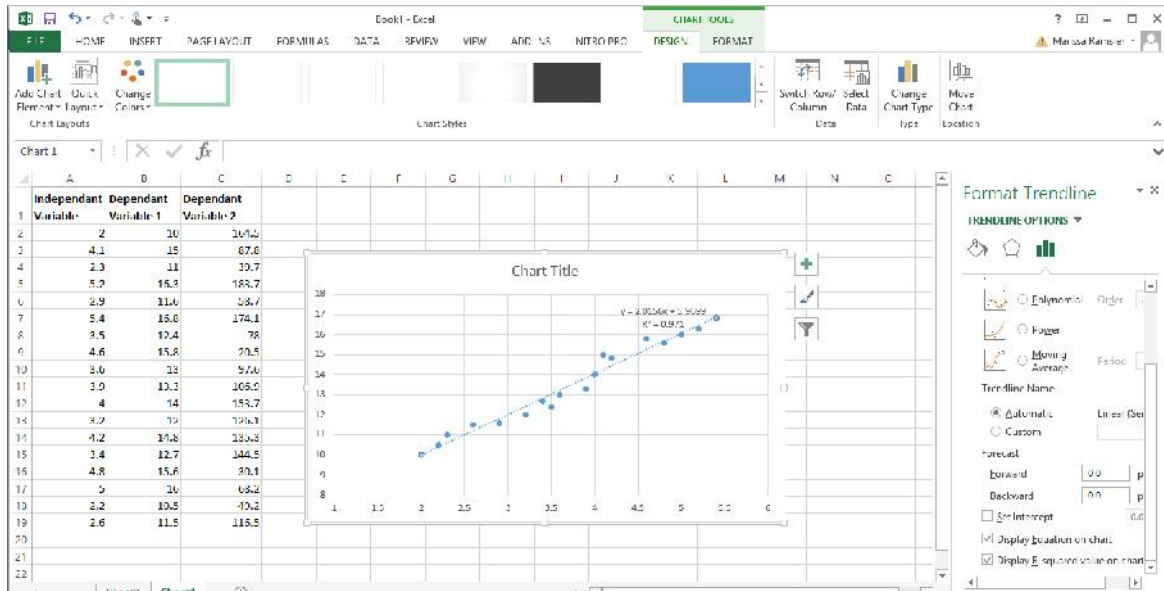
(12) Adjust your chart axes – double click on the numbers in the axes, and a dialog box should pop up on the right. Change the min and max values to bracket your data, then close the dialog box.



(13) Right click on the data points (the actual dots in the chart) and “Add Trendline”



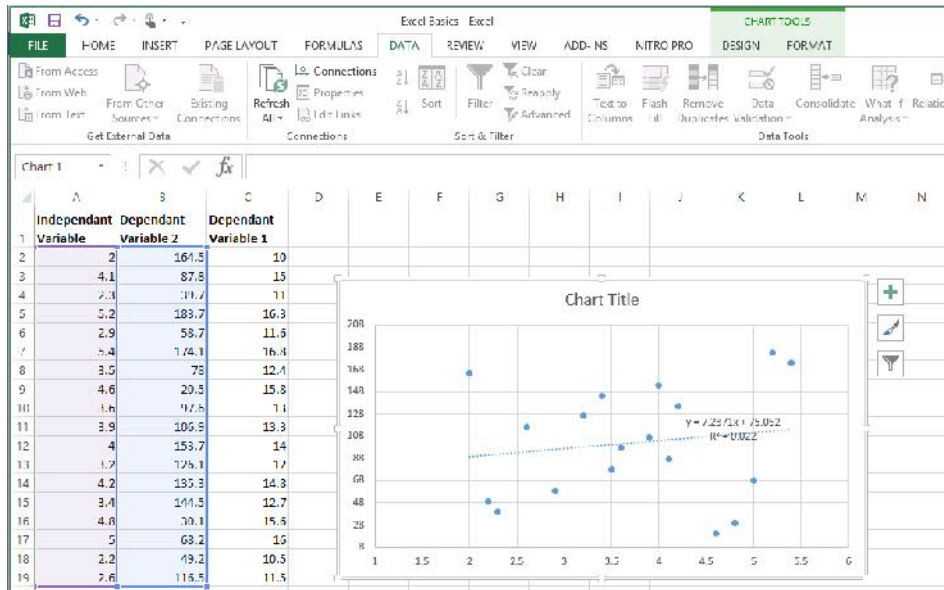
(14) In the trendline dialog box that pops up on the right, select linear (if your data are linear), and scroll down to check the boxes for “Display equation...” and “Display R-squared...”, then close the dialog box.



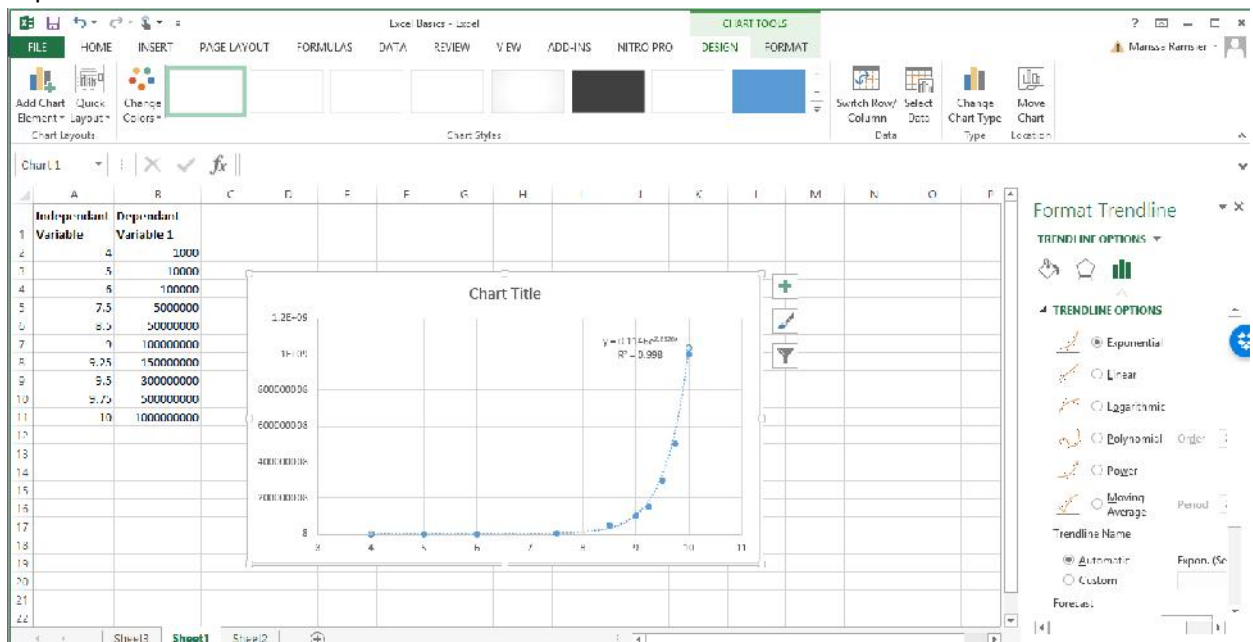
(15) The equation in the trendline should match the equation achieved with your OLS. If not, check to make sure you have your columns organized so that the independent variable is to the left of the dependent, then redo the chart. Note: If you change any of the data, you will need to delete and re-add the trendline, as the equation of the line and the  $R^2$  will not automatically update.

(16) Now, move your columns around so that your next dependent variable is next to the independent variable, and repeat all steps. For this analysis (see below), the relationship was not significant ( $R^2=0.02$ , ANOVA  $F=0.36$ ,  $P=0.56$ ). Note that since the  $P$  was not  $<0.01$ , I am going to report the actual value, which you would do for any value  $>0.01$ , such as 0.02, 0.04, 0.12...).

SUMMARY OUTPUT								
<b>Regression Statistics</b>								
Multiple R	0.14849739							
R Square	0.02204999							
Adjusted R Square	-0.039071886							
Standard Error	52.13094022							
Observations	18							
<b>ANOVA</b>								
	df	SS	MS	F	Significance F			
Regression	1	987.7377917	987.7377917	0.36075447	0.55660073			
Residual	16	43497.1372	2718.5742					
Total	17	44477.925						
<b>Coefficients</b>								
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	75.05205162	46.43866944	1.615154221	0.125804522	-23.3935298	173.4976331	-23.3935298	173.4976331
X Variable 1	7.237116155	12.04924076	0.600628396	0.55550073	-18.30613310	32.7803655	-18.30613310	32.7803655

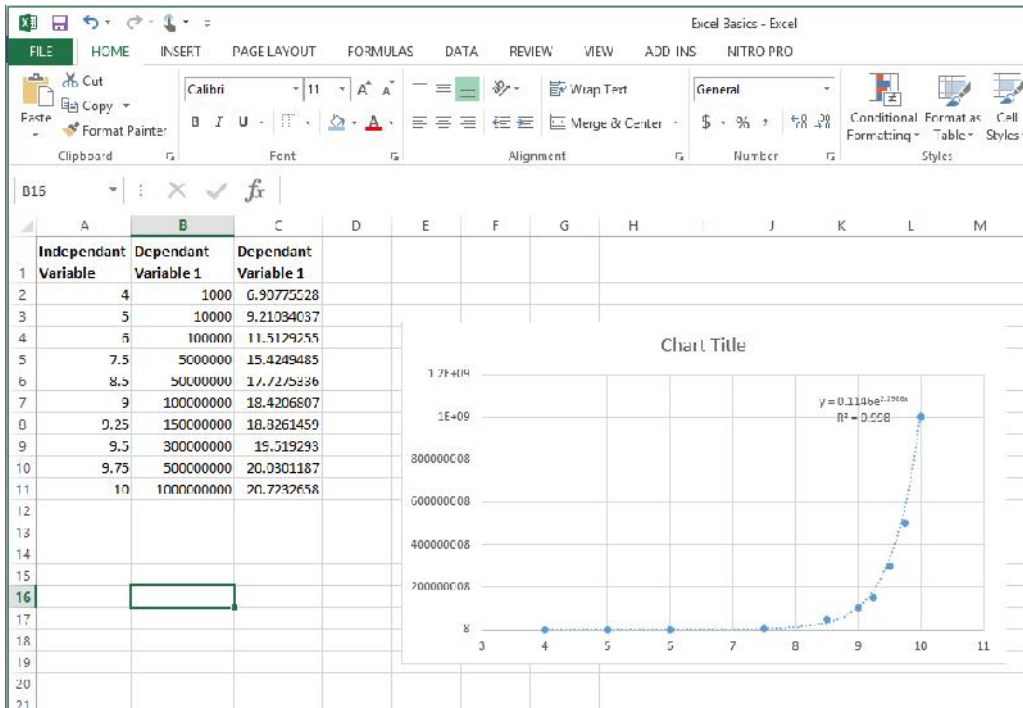
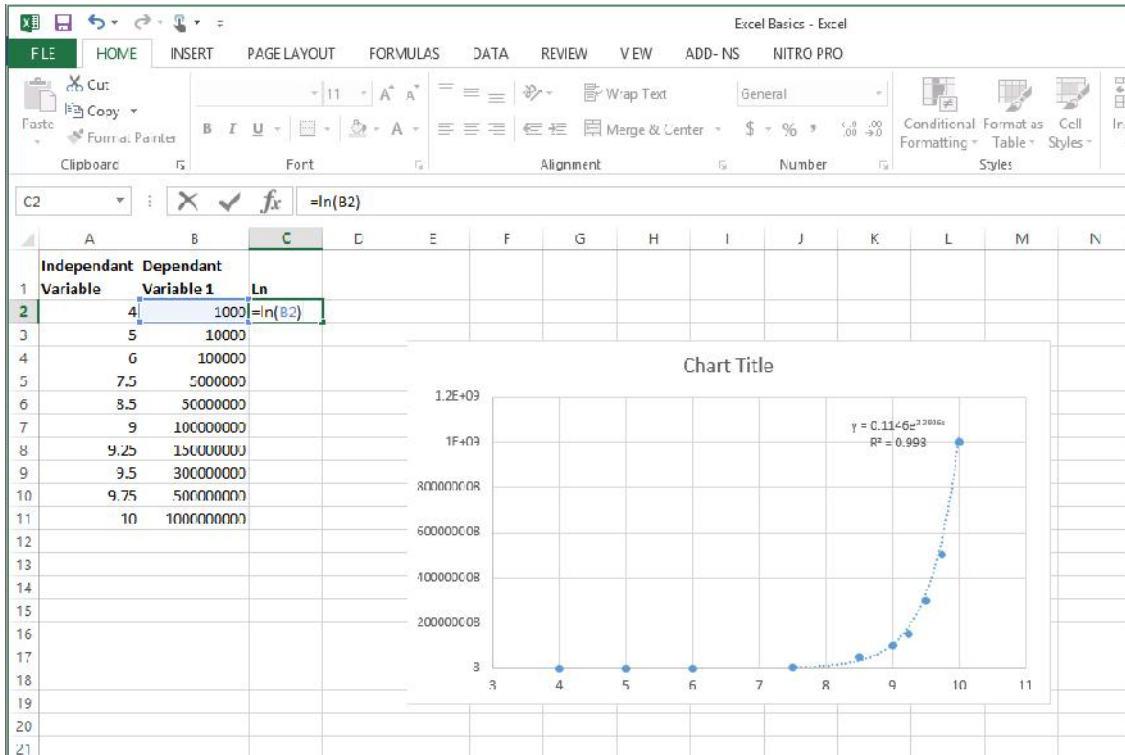


(17) There are various other things that can happen with your data. For example, there may be an exponential relationship between the data (common in growth studies), in which case an exponential trend line would need to be fit:





(18) You cannot run an OLS with exponential data and get a correct P value for the relationship, unless you first log transform the data. For example:



Select the Ln (Natural log) transformed data for your Y variable:

The screenshot shows the 'Regression' dialog box in Excel. The 'Input Y Range' is set to '\$C\$2:\$C\$11' and the 'Input X Range' is '\$A\$2:\$A\$11'. The 'Confidence Level' is 95%. Under 'Output options', 'New Worksheet Ply' is selected. Under 'Residuals', 'Residuals' and 'Standardized Residuals' are checked. Under 'Normal Probability', 'Normal Probability Plots' is checked. A scatterplot is visible in the background with a regression line.

Now, you can use the OLS output R Square and ANOVA F and P (significance), but use the equation of the line from the scatterplot of the non-log transformed data.

The screenshot shows the regression output table in Excel. The table includes 'Regression Statistics' and 'ANOVA' sections. The R Square value is 0.998024741. The ANOVA table shows a significant F-statistic of 4042.10129 with a p-value of 4.1658E-12.

Regression Statistics	
Multiple R	0.999011882
R Square	0.998024741
Adjusted R Square	0.997777833
Standard Error	0.231317858
Observations	10

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	216.2845599	216.2845599	4042.10129	4.1658E-12
Residual	8	0.428063612	0.053507952		
Total	9	216.7126235			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-2.166590655	0.292368672	-7.410474737	7.54359E-05	-2.840794021	-1.492387289	-2.840794021	-1.492387289
X Variable 1	2.292507628	0.036059877	63.5775219	4.1658E-12	2.209443403	2.375751854	2.205443403	2.375751854