Simple Linear Regression Excel 2010 Tutorial

This tutorial combines information on how to obtain regression output for Simple Linear Regression from Excel and some aspects of understanding what the output is telling you. Most interpretation of the output will be addressed in class.

The scenarios for this (and all of the Excel Regression tutorials) are described in the Regression Scenarios Word file at: http://faculty.ung.edu/kmelton/Documents/RegressionScenarios.docx.

The Reg1 Excel file for this tutorial is located at http://faculty.ung.edu/kmelton/Data/Reg1.xlsx. The Excel file for this tutorial contains data on five sheets accessed at the bottom left of the page. One tab is related to each of the scenarios described in the Word document. [Note: The data used to produce the output on the fifth tab is not the data that will be used for this scenario in the other Excel Regression Tutorials.]

Obtaining Simple Linear Regression (SLR) Output

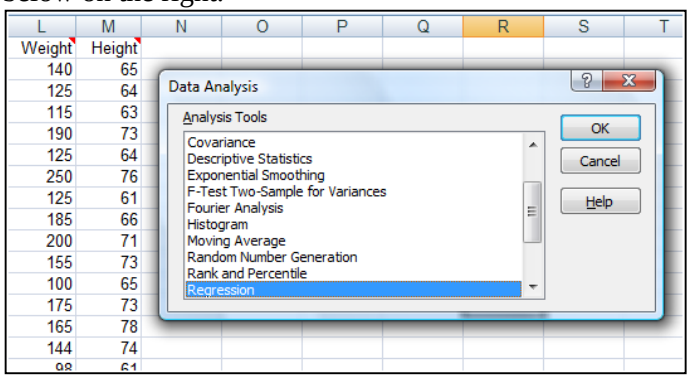
- Start with your model statement (based on the theory to be tested). This will identify the variables.
  - Rewrite the Simple Linear Regression model statement ($Y = \beta_0 + \beta_1 X + \varepsilon$) using variable names from the problem (e.g., Weight = $\beta_0 + \beta_1$Height + ε)
- Recognize the way Excel wants the data to be displayed in the spreadsheet
  - One column of data for each variable with the name of the variable in the first row
  - For SLR the order of the variables does not matter; but as we move forward to multiple regression, having the dependent variable (Y) first is easier.
- Enter (or confirm) data in the needed format
- Use the Regression procedure in the Data Analysis Tools of Excel to obtain the output
  - Be careful, Excel asked you to identify Y first and then X
  - Be sure to select your variable names along with the data and tell Excel that you have the labels
  - Do not select the other options in the Input section of the dialogue box
  - By selecting Output Range and a cell for the upper left corner of your output, you can have the output placed on the same page with your data.
  - All other choices below there are optional (and depend on what additional output that you want Excel to provide). We won't use the Standardized Residuals, the Line Fit Plots, or the Normal Probability Plots for this course.
- Clean up the output
  - Remove unnecessary parts of the output
  - If you are going to print the output, position the output so that all output from the same model statement prints together. Do not split the first three sections across different pages (Summary Output, ANOVA, and Coefficients).
- Move on to the hard part…understanding what the output tells you.

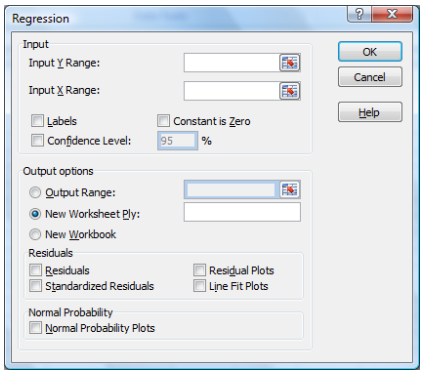
Example 1: Using the Weight Scenario consider the analysis that would be needed to address either of two theories. In both of these we are trying to predict Weight. Theory 1: Height can be used as a predictor for weight? Theory 2: Taller people would be expected to weigh more (than shorter people)? For either of these, if we believe there is a straight line relationship between height and weight, we would the use the model statement: Weight = $\beta_0 + \beta_1$Height + ε. Based on this we can see that we will need two columns of data—one with the weight (as Y) and one with the height (as X) for the individuals in the data set.

When you look at the Weight tab of the Excel file, you will see that the data include these two variables and some more. In addition, you will see a scatter diagram so you can visualize the relationship between the variables. This is not necessary for regression analysis (and will not even be an option when you have multiple independent variables.

We are not concerned with the additional variables (at this point). The variables we need are Height in column B and Weight in column C. Although using the data from these locations would work fine for Simple Linear Regression, copying the data to a new location in a format that reduces the likelihood that we select the wrong variables for analysis is a good idea. Since Excel will ask for the dependent variable (Y) first and then the independent variables (Xs), let's copy the data for Weight into column L and the data for Height into column M.

Obtaining Regression Output: Next we select Data Analysis from the Data tab and scroll down to select Regression. [If you don't see Data Analysis and you are using Excel for Windows, you will need to add this option. Instructions were in the first Excel Tutorial for the semester.] The resulting dialogue box is shown below on the right.





To complete the dialogue box, you must understand Excel's language.

- Input Y Range requires you to identify the data for the dependent variable. To do this click in the white box to the right of "Input Y Range:" and then drag over the values of Weight including the variable name in the first row.
- Input X Range requires you to identify the data for the independent variable. Repeat the process you used for the Y values except selecting the Height data and variable name.
- Labels is asking you if you the first row of the data you selected contains the names for the variables. Since you were instructed to select the variable names, click on the little square to indicate that you did select the labels. Selecting the variables is a good idea if you want to be able to match your output to the data and (as we move forward to multiple regression) if you want to know what variable is related to each row in the coefficients section of the output. [Warning: If you did not select the variable names and you do click the box next to Labels, Excel will use the first row of your data as a variable name and will not include the value in the calculations…so you will get wrong output!]
- Constant is Zero is asking if you want to force the fitted line to go through the origin. This is not recommended for most cases (and will not be used for any analysis in this course).
- Confidence Level allows you to have Excel complete confidence intervals for β_0 and β_1. We will not use these functions in this course.
- Output options tells Excel where you want to put your output.
    - Output Range allows you to have the output put on the same sheet as your data. This is the approach that will be shown in the tutorial. To put the output on the same page as the data, click on the write box to the right of "Output Range:" and then click on the cell that will be the upper left cell used in the output. For this tutorial, put the output in cell O1. NOTE: If Excel is about to over-write other data, you will receive a warning.
    - New Worksheet Ply allows you to name a new sheet within the same file for your output.
    - New Workbook allows you to save the output to a totally new file.

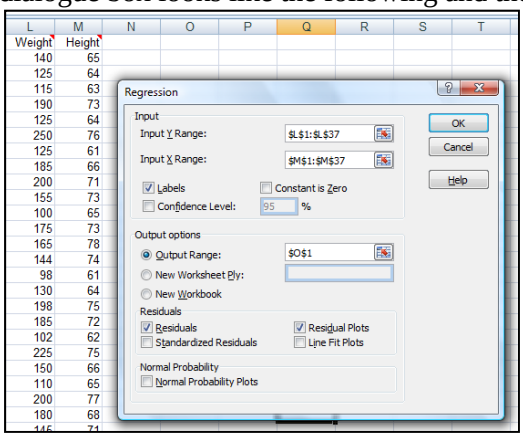
If you click OK at this point, you will receive standard regression output consisting of three parts (Summary Output, ANOVA table, and the coefficients section). Everything beyond this is considered optional output.

- Residuals

- Residuals creates one row of data for observation in the data set. Excel uses the fitted equation to estimate a value of Y for the observed value of X and compares the estimated Y to the actual Y.
- Residual Plots creates a scatter diagram showing the values of X on the horizontal axis and the Residuals (the actual Y value – the estimated Y value) on the Y axis. This allows you to see if there are patterns in "errors."
- Standardized Residuals uses a transformation similar to finding a Z score to determine the number of standard deviations each observation is from the line. This allows you to check for extreme values (that may be a signal of incorrect data or outliers).
- Line Fit Plots shows a "not too good" graph of what the line looks like on a scatter diagram created using the X and Y in the model statement.

- Normal Probability Plots provides a way to see if the distribution of your Y variable is approximately normal. We will not use this function in Excel.

Check to see that your completed dialogue box looks like the following and then click OK.



The standard output obtained from selecting the Y values (weights in this example), the X values (heights in this example), and indicating where to put the output give the following.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.770399
R Square	0.593514
Adjusted F	0.581559
Standard E	24.03504
Observatio	36

ANOVA

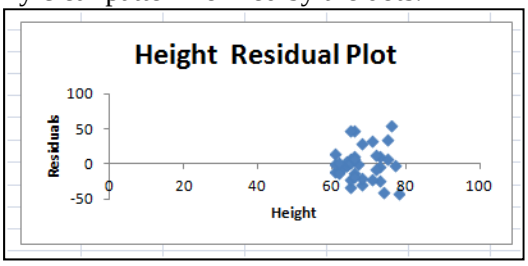
	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	1	28678.4	28678.4	49.64381	3.91E-08
Residual	34	19641.24	577.6834		
Total	35	48319.64			

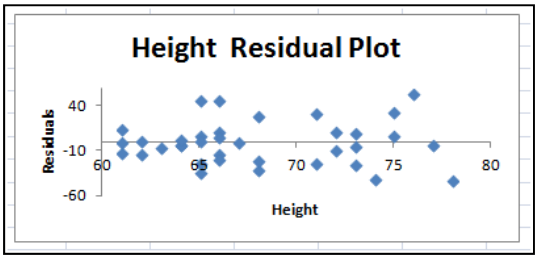
	Coefficients	*tandard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-241.917	55.81125	-4.33456	0.000123	-355.339	-128.495	-355.339	-128.495
Height	5.767981	0.818637	7.045836	3.91E-08	4.104311	7.431652	4.104311	7.431652

Checking the Residuals box provides the following output (only the first 7 of the 36 rows is shown here):

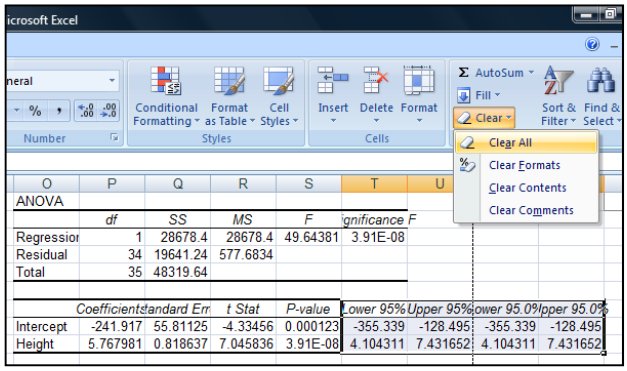
RESIDUAL OUTPUT		
Observation	*dicted Wei*	*Residuals*
1	133.0016	6.998389
2	127.2336	-2.23363
3	121.4656	-6.46565
4	179.1455	10.85454
5	127.2336	-2.23363
6	196.4494	53.55059
7	109.9297	15.07031
8	138.7696	46.23041

Checking the Residual Plots box provides the output in the scatter diagram on the left below. The diagram on the right is the same data with the horizontal and vertical axes adjusted to focus on the data. When the assumptions behind simple linear regression are met, we would expect to see the points plot in a horizontal band without any clear pattern formed by the dots.





Cleaning up the Output: Since we will not be using the last four columns in the coefficient section, we can get rid of them. Highlight the three rows of the four columns (as shown in the figure on the left below) and then from the Home tab at the top of the screen, select Clear All from the drop down under "Clear." Some other clean up that makes reading the output a little easier is to resize the columns, center the values in the df column of the ANOVA table, and shorten the Significance F label to Sig. F. These are shown in the figure on the right below.



SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.770398888				
R Square	0.593514446				
Adjusted R Square	0.581558989				
Standard Error	24.03504499				
Observations	36				
ANOVA					
	df	*SS*	*MS*	*F*	*Sig. F*
Regression	1	28678.40371	28678.4	49.64381	3.91E-08
Residual	34	19641.23518	577.6834		
Total	35	48319.63889			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-241.9171823	55.81125272	-4.33456	0.000123	
Height	5.767981439	0.818636881	7.045836	3.91E-08	

Each of the following contains a signal that you made a mistake. Can you spot the problem in each?

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.770487				
R Square	0.59365				
Adjusted F	0.581336				
Standard E	24.36485				
Observatio	35				
ANOVA					
	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	28620.09	28620.09	48.21071	6.19E-08
Residual	33	19590.31	593.6458		
Total	34	48210.4			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	
Intercept	-243.841	56.95724	-4.28113	0.00015	
65	5.793306	0.834362	6.943394	6.19E-08	

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.770399				
R Square	0.593514				
Adjusted R	0.581559				
Standard E	3.210234				
Observatior	36				
ANOVA					
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	511.6095	511.6095	49.64381	3.91E-08
Residual	34	350.3905	10.3056		
Total	35	862			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	
Intercept	52.53384	2.259344	23.25182	1.87E-22	
Weight	0.102898	0.014604	7.045836	3.91E-08	

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.7704				
R Square	0.59351				
Adjusted R S	0.58156				
Standard Err	24.035				
Observations	36				
ANOVA					
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	28678.4	28678.4	49.64381	3.91E-08
Residual	34	19641.24	577.6834		
Total	35	48319.64			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	
Intercept	-241.917	55.81125	-4.33456	0.000123	
X Variable 1	5.76798	0.818637	7.045836	3.91E-08	

In the output on the left, there are only 35 observations and the last row of the Coefficients section says 65. Looking back at the data, we see that the first observations of Height was 65. This is a signal that the dialogue box was completed incorrectly—the data without the variable names were selected and the Labels box was checked. Excel used the first row of data as variable names!

In the output in the center, part of the Summary Output and part of the ANOVA section match the output that we received, but most of the numbers differ. Looking at the Coefficients section, we see that the last row is labeled "Weight." Since this row show relate to the coefficient of our X variable, we would expect to see Height listed here. Again, this is a signal that the dialogue box was completed incorrectly—in this case, the data for the Y values and the data for the X values were swapped.

In the output on the right, all of the numbers match our output. The only difference is the label in the last row of the Coefficients section of "X variable 1." In this case, the data were selected for analysis without selecting the variable names and the box next to Labels was not checked. Although this is not wrong, this makes it difficult to confirm that we have selected the variables correctly and will create lots of confusion when we get to multiple regression where we will end up with multiple X variables.

Other Examples

Example 2: Predicting sales using the dollars put into advertising. ["Sales" Worksheet accessed at the lower left of the same file.] Note: Data were collected on advertising expenses and on square feet of shelf space. This example focuses on the use of advertising expenses as a possible predictor of sales.

Using the Model statement: Sales = $\beta_0 + \beta_1$AdExp + ε and following the same steps as the previous (this time selecting the Sales data for Y and the AdExp data for X), we obtain the following output and residual plot. The axes on the residual plot have already been adjusted to focus on the data. In this case, we see that as advertising expenses increase, there is more variation between our expected and observed sales. One of the assumptions behind regression is that the variation should remain constant across the horizontal axis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.766663
R Square	0.587772
Adjusted R	0.564871
Standard E	776.8759
Observatio	20

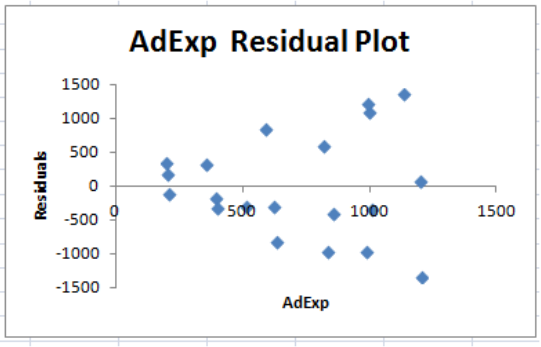
ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	15489879	15489879	25.6652	8.05E-05
Residual	18	10863652	603536.2		
Total	19	26353531			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	1133.23	409.9154	2.764546	0.01277
AdExp	2.652429	0.523566	5.066084	8.05E-05

RESIDUAL OUTPUT

Observatior	*edicted Sal*	*Residuals*
1	1666.368	343.6316
2	1676.978	173.0219



Example 3: Predicting demand for electricity (Load) using the predicted high temperature for the day. ["Power" Worksheet accessed at the lower left of the same file.]

Using the Model statement: Load = $\beta_0 + \beta_1$Temp + ε and following the same steps as the previous (this time selecting the Load data for Y and the Temp data for X), we obtain the following output and residual plot. The axes on the residual plot have already been adjusted to focus on the data. In this case, we see a clear pattern in the residuals as we move across the horizontal axis. There is a U-shaped pattern (a curve). Again, this is inconsistent with what we would expect. A curve on the residual plot is a signal that the assumption that the relationship between the variable s follows a straight line may not be appropriate. We may need to consider a model that allows for a curve.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.918339
R Square	0.843347
Adjusted F	0.836536
Standard E	10.32357
Observatio	25

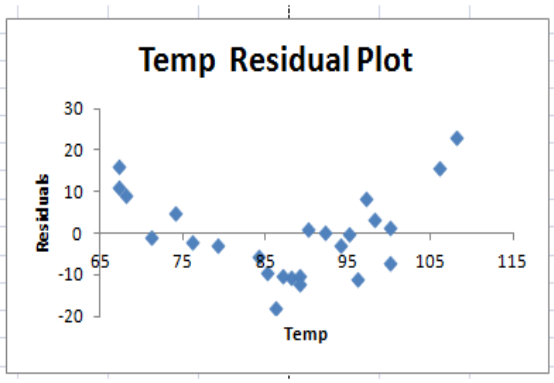
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	1	13196.4	13196.4	123.8214	9.82E-11
Residual	23	2451.25	106.5761		
Total	24	15647.65			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	-47.3935	15.66766	-3.02493	0.006027
Temp	1.976458	0.177619	11.12751	9.82E-11

RESIDUAL OUTPUT

Observatior	*redicted Loa*	*Residuals*
1	138.3936	-2.39357
2	142.3465	-10.6465



Moving on to the hard part…Understanding what the output tells us

Example 4: Predicting Time to Relief for a medicine (Time) using the predicted the age (Age) of the individual.

Using the Model statement: Time = $\beta_0 + \beta_1$Age + ε to direct our actions, requires us to think about how to put the data into Excel. The original data were provided as follows:

Liquid			Pill			Shot	
Age	Time		Age	Time		Age	Time
51	22		46	28		37	19
36	25		40	24		60	17
31	20		26	23		25	21
20	25		32	25		38	20
20	24		62	30		22	20
35	23		44	27		29	20
50	22		32	24		41	19
43	22		23	22		59	17

To use Excel to analyze the data requires us to adhere to the expectations that each variable is provide in a single column. Therefore, we need one column for Time (our Y variable) and one column for Age (our X variable). The data have been reorganized on the "Time" Worksheet accessed at the lower left of the same file.

Once this is done, we follow the same steps as the previous (this time selecting the Time data for Y and the Age data for X). We obtain the following output and residual plot.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.006814
R Square	4.64E-05
Adjusted F	-0.04541
Standard E	3.343562
Observatio	24

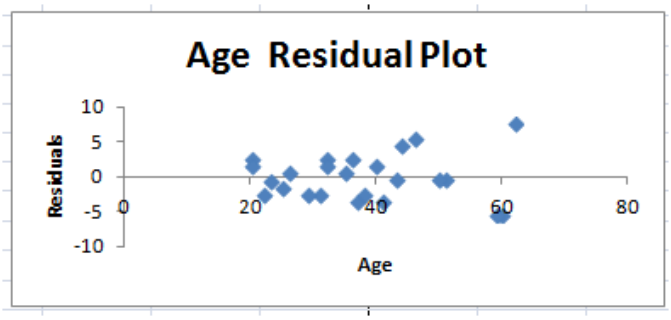
ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	0.011419	0.011419	0.001021	0.974793
Residual	22	245.9469	11.17941		
Total	23	245.9583			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	22.52521	2.201158	10.23335	7.92E-10
Age	-0.00178	0.055681	-0.03196	0.974793

RESIDUAL OUTPUT

Observatior	*redicted Tin*	*Residuals*
1	22.43446	-0.43446
2	22.46115	2.538849
3	22.47005	-2.47005



This time, we don't see major issues related to the residuals. That does not mean that we are ready to predict time to relief using the age of the individual. If the only goal of simple linear regression had been to fit a straight line to our sample data, we could have done this from the scatter diagram.

Or goal was to use data from a sample to draw a conclusion about whether the same relationships would be true in the population that produced the sample. Understanding the numbers on the output help us determine if we can generalize from the current sample to the larger population. Like our previous hypothesis test, the data in our output answers the question, how unusual would it be for us to see something this extreme if there were no linear relationship between the variables? Specifically, the Significance F (a special form of a p value) and the p value in the row that corresponds to Age are assessing if the data supports a conclusion that people of different ages would be expected to see different times to relief. A low p value would indicate that it would be highly unlikely to see something this extreme; and a high p value would prevent us from concluding that there was enough evidence to support that a linear relationship exists. In this case, the p value is .974793 (very high given that p values must be between 0 and 1)! NOTE: This does not mean that age has no relationship with (role in predicting) time to relief; but it does appear that trying to use a straight line relationship where age is the only predictor is not supported.

The last sheet on the file: The "Final" worksheet accessed at the lower left provides output from the model: Final = $\beta_0 + \beta_1$Midterm + ε where Midterm is a student's midterm grade and Final is that same student's final grade in a course. You will see small triangles in the upper right corner of most cells in the Excel regression output. If you move your mouse across the cell, notes will appear describing what the entry in the cell represents (or how it relates to other values in the output). This page is provided as a reminder of what the entries represent.

Remember, Regression output provides answers to lots of questions—we just need the equivalent of Alex Trebek's notes to know the questions. (For those who are not familiar with the name--Alex Trebek is the host of Jeopardy on TV. Jeopardy is a game show where contestants are given the answers and they most provide the questions.)