Multiple Linear Regression Excel 2010 Tutorial
For use with more than one quantitative independent variable

This tutorial combines information on how to obtain regression output for Multiple Linear Regression from Excel (when all of the variables are quantitative) and some aspects of understanding what the output is telling you. Most interpretation of the output will be addressed in class. This tutorial assumes that you have already been through the one for Simple Linear Regression.

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 Reg2 Excel file for this tutorial is located at http://faculty.ung.edu/kmelton/Data/Reg2.xlsx. The Excel file for this tutorial contains data on three sheets accessed at the bottom left of the page. Each tab is related to one of the scenarios described in the Word document. [Note: The data used on the tab labeled "Final" is not the same data that was used for the last tab in Simple Linear Regression.]

Obtaining Multiple Linear Regression Output

- Start with your model statement (based on the theory to be tested). This will identify the variables.
  - Rewrite the Multiple Linear Regression model statement ($Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \ldots \beta_kX_k + \ + \varepsilon$) using variable names from the problem (e.g., $Sales = \beta_0 + \beta_1AdExp + \beta_1ShelfSpc + \varepsilon$)
- 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 Multiple Regression the order of the variables does matter; all of the independent variables (Xs) must be in consecutive columns, and having the dependent variable (Y) first is easier. The order of the independent variables is not important—just that they are in consecutive columns. The coefficient section of your output will have one row for the intercept and one row for each independent variable (in the order of the columns). If you run the analysis twice with the only change being the order of the columns for the independent variable, the only change in the output will be the order of the rows in the coefficient section.
- 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 (dragging over multiple columns so that you include all of the X variables)
  - 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). For the most part, we will not use output beyond the basic output for what we are covering in this course—so you can leave the boxes next to the other choices blank for multiple regression.
- 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 predicting sales example, consider the analysis that would be needed to address the following theories/questions:

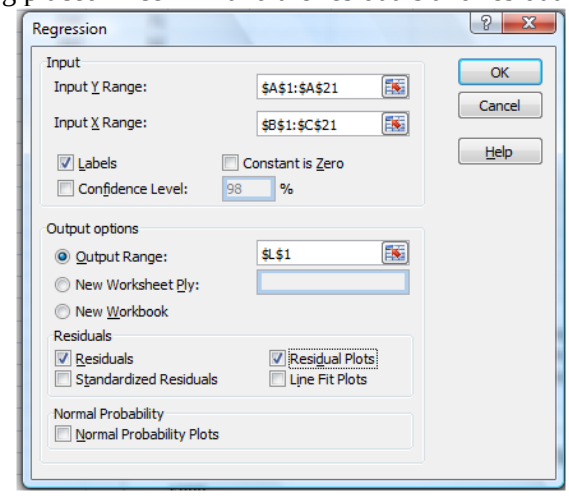
- Should we use advertising expenses and the amount of visibility of the product in the store to estimate sales? [And, would this be better than our previous analysis where we only considered advertising expenditures to help estimate sales?]
  - When we increase the shelf space (without changing the money we put into advertising), can we estimate the impact of additional space on sales?
  - When we increase advertising expenses (without changing the amount of shelf space for the product), can we estimate the impact of additional advertising expenditures on sales?

All of these theories can be evaluated with the same model statement:

$$\text{Sales} = \beta_0 + \beta_1 \text{AdExp} + \beta_2 \text{ShelfSpc} + \varepsilon$$

If we think about how this equation would look on a graph, we can visualize this in two dimensions by asking what happens when we how the value of one of the independent variables constant? For example, if we were to keep the amount of shelf space the same (say 30 sq. ft) and only allow the amount spent on advertising to vary, β_2ShelfSpc would be a constant; and we would have a straight line with slope β_1 and intercept ($\beta_0 + \beta_2$ShelfSpc). If we looked at what would happen when we held shelf space constant at some other value (say 50 sq. ft.), the result would still be a line with slope β_1, but the intercept would move. This means the lines would be parallel to each other.

To obtain estimates for the βs in the model and to assess if there is sufficient evidence to generalize from our sample data to the larger population, we need to obtain the regression output. The data are on the Sales tab (accessed at the lower left) of the Excel file. We see that the data are already entered with one column for each variable, and the two independent variables are in consecutive columns. The method to obtain output for this model is very similar to the method we used in Simple Linear Regression. We use the same regression procedure in the Data Analysis Tools. Like before, we use the model statement to help us fill in the dialogue box. Since Sales is our Y variable, we drag over the data for Sales (including the variable name) to fill in the first box. Now we have two X variables. Therefore, we drag over the data and variable names for both variables (AdExp and ShelfSpc) to identify the X data. Be sure to select the Labels box indicating that you did select the variable names. The illustration below shows the upper left corner of the output being placed in cell L1 and the residuals and residual plots being requested.



Once you click OK and clean up the output, you get the following output. Note that this time you have three rows in the coefficient section. In addition, there are two residual plots—one for each independent variable. Also note that the Significance F value in the ANOVA table no longer matches any of the p

values in the coefficient section. [I guess you could say that we have moved to the Double Jeopardy round of answers and questions!] Since the objective of this tutorial is learning how to obtain output, the interpretation of the output will be addressed in class.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.949276
R Square	0.901125
Adjusted R Squ	0.889492
Standard Error	391.5064
Observations	20

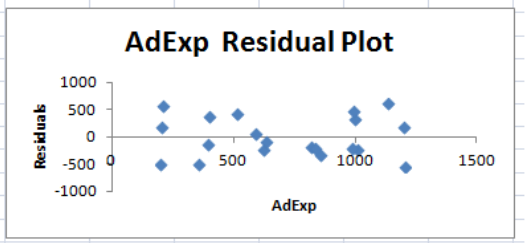
ANOVA

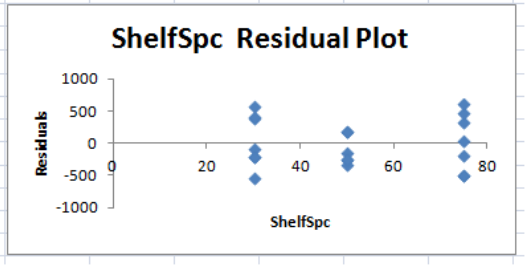
	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	23747818	11873909	77.46687	2.87E-09
Residual	17	2605713	153277.3		
Total	19	26353531			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	-558.546	309.5127	-1.8046	0.088883
AdExp	2.548136	0.264233	9.643512	2.63E-08
ShelfSpc	34.1205	4.648559	7.340016	1.15E-06

RESIDUAL OUTPUT

Observation	*edicted Sal*	*Residuals*
1	2512.666	-502.666
2	1669.846	180.1536
3	2905.079	-505.079
4	995.0809	579.9191
5	3503.891	46.10886
6	2159.088	-144.088
7	4089.962	-181.962
8	1484.323	385.677
9	4540.982	336.0175
10	1777.359	412.6414
11	4538.434	466.5656
12	2740.063	-240.063
13	3338.875	-333.875
14	3726.192	-246.192
15	4892.625	607.3747
16	2083.135	-88.1349
17	2597.858	-207.858
18	4205.242	184.7583
19	2987.723	-202.723
20	3535.572	-546.572





Other Examples:

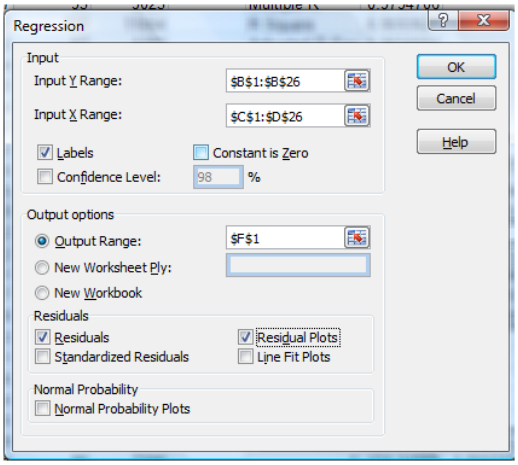
Example 2: Predicting demand for electricity ["Power" Worksheet accessed at the lower left of the same file.] In the Simple Linear Regression tutorial we looked at using the forecasted high temperature (Temp) for the day as a predictor of the demand (Load) for electricity. The residuals from this situation showed a clear pattern—one that suggested that the relationship may not follow a straight line. There are many ways to introduce a curve into the equation; but one of the simplest is to include a quadratic term (X^2 term). Therefore, we will consider the following model: $\text{Load} = \beta_0 + \beta_1\text{Temp} + \beta_2\text{Temp}^2 + \varepsilon$.

To get output for this model, we will need a column of data for each of the variables (Load, Temp, and Temp^2), we will need to create data for the Temp^2 term, and the Temp and Temp^2 data will need to be in consecutive columns. As you can see in the following screen image, the Temp data were moved to column C and a Temp(sq) variable has been entered in column D. To create the Temp(sq) term, click on cell D2 and enter either one of the following formulas; =C2^2 [this one raises the value in cell C2 to the 2^{nd} power] or =C2*C2 [this one multiples the value in C2 by itself]. Then copy/fill the rest of the column with the formula. Once this is done, you are ready to generate the Excel output. As usual, the model

statement tells you the data to select (Load for the Y values and both Temp and Temp(sq) for the X values). The output is shown below.

=C2^2

	B	C	D
	Load	Temp	Temp(sq)
	136	94	8836
	131.7	96	9216
	140.7	95	9025
	189.3	108	11664
	96.5	67	4489
	116.4	88	7744
	118.5	89	7921
	113.4	84	7056
	132	90	8100
	178.2	106	11236
	101.6	67	4489
	92.5	71	5041
	151.9	100	10000
	106.2	79	6241



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9794706
R Square	0.9593627
Adjusted R Squ	0.9556684
Standard Error	5.3762035
Observations	25

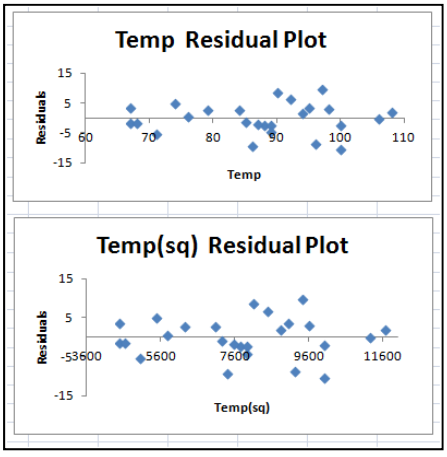
ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	15011.77	7505.886	259.6872	4.99E-16
Residual	22	635.8784	28.90356		
Total	24	15647.65			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	385.04809	55.17244	6.978994	5.27E-07
Temp	-8.292527	1.299045	-6.38356	2.01E-06
Temp(sq)	0.0598234	0.007549	7.925143	6.9E-08

RESIDUAL OUTPUT

Observation	*redicted Loa*	*Residuals*
1	134.14986	1.850143
2	140.29768	-8.59768
3	137.16395	3.536054
4	187.23497	2.065033
5	97.995898	-1.4959
6	118.5779	-2.1779
7	120.87411	-2.37411
8	110.58953	2.810471
9	123.28996	8.710035
10	178.21562	-0.01562
11	97.995898	3.604102
12	97.84829	-5.34829
13	154.0291	-2.1291
14	103.29612	2.903882
15	143.55107	9.648934

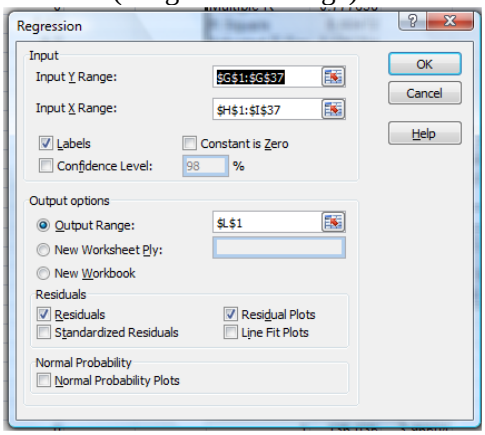


Example 3: Predicting weight using height and change in someone's pocket ["Weight" Worksheet accessed at the lower left of the same file.] This example is a little "far-fetched," but allows us to see how the regression output signals us that a theory may not be quite right. This example starts off like our first example in this tutorial and considers whether both variables are helping us predict the dependent variable. The model statement is set up in the same format as in our first example:

$$\text{Weight} = \beta_0 + \beta_1\text{Height} + \beta_2\text{Change} + \varepsilon$$

As you can see from the illustration below, the data have been moved to columns G, H, and I—with the Weight data (the Y variable) in G and the two X variables (Height and Change) in the next two columns.

G	H	I
Weight	Height	Change
140	65	0.23
125	64	0.9
115	63	0
190	73	0.54
125	64	0.7
250	76	0.2
125	61	0.1
185	66	0.25
200	71	0.25
155	73	1.05
100	65	0.85
175	73	0.71
165	78	0.27
144	74	0.1
98	61	0.4
130	64	0.77
198	75	1.06



Selecting this data in the dialogue box and putting the output in L1 provides the results shown below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.777638
R Square	0.60472
Adjusted R Squ	0.580764
Standard Error	24.05787
Observations	36

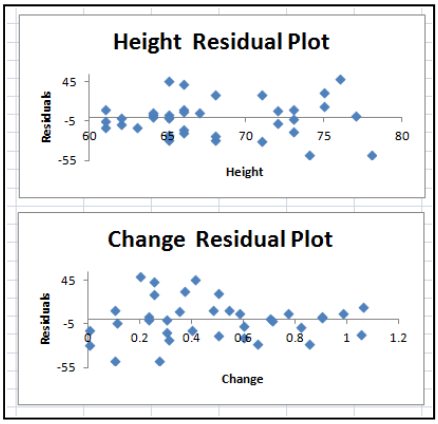
ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	29219.86	14609.93	25.24259	2.23E-07
Residual	33	19099.77	578.781		
Total	35	48319.64			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	-242.463	55.8671	-4.34	0.000127
Height	5.869842	0.826154	7.105022	3.89E-08
Change	-13.2235	13.67161	-0.96722	0.340468

RESIDUAL OUTPUT

Observation	*dicted Wei*	*Residuals*
1	136.035	3.96504
2	121.3054	3.694621
3	127.3367	-12.3367
4	178.8944	11.10558
5	123.9501	1.049923
6	200.9999	49.00007
7	114.2746	10.72535
8	141.6403	43.35967
9	170.9895	29.01046
10	172.1504	-17.1504
11	127.8364	-27.8364
12	176.6464	-1.64642
13	211.814	-46.814
14	190.5826	-46.5826
15	110.3076	12.3076



When we look at the standard regression output and focus on the Significance F and the p values in the coefficient section, we see that the p value in the "Change" row is higher than we would expect. This leads us to the need to understand what this might be telling us! Good discussion for class!