Multiple Linear Regression Excel 2010 Tutorial
For use when interaction is considered

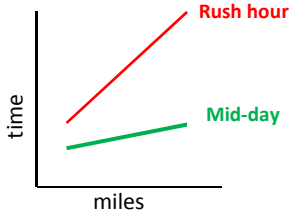
This tutorial combines information on how to obtain regression output for Multiple Regression from Excel (when all of the variables interaction is a possibility) and some aspects of understanding what the output is telling you. Most interpretation of the output will be addressed in class. This tutorial assumes that you have been through the Multiple Regression tutorials covering the basics and covering including qualitative variables.

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 Reg4 Excel file for this tutorial is located at http://faculty.ung.edu/kmelton/Data/Reg4.xlsx. The Excel file for this tutorial contains data on four 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.]

Recognizing Interaction

In the previous two tutorials on Multiple Regression there was an implied assumption that you could talk about the impact of changing the value of one of the independent variables without knowing the value of other independent variables in the model (i.e., you could make statements like "holding everything else constant" or "while controlling for…"). Few things in life happen independently of other things. When you find yourself saying, "(paraphrased) The impact of changing X (one predictor variable) depends on the value of another predictor variable," you are suggesting that there is interaction between the two independent variables. For example, suppose you were trying to estimate the time to get to the airport based on time of day and miles to the airport. If someone were to ask "How much longer will it take to get to the airport as my distance from the airport increases?" people in large cities would probably agree that the response would be, "That depends on the time of day you plan to travel." Graphically, this would show up as lines that are not parallel to each other. In the "time to get to the airport" example, the following might represent the relationship between distance and time.



During the middle of the day, travel moves at close to speed limit. During rush hour, traffic grinds to a snail's pace – so the expected time to travel each mile increases. Although people travelling a short distance may not experience much difference in travel time based on the time of day, people with longer distances to travel will take considerably longer.

Writing Model Statements that Allow for Interaction

Sometimes, interaction is referred to as a cross-product term. This name for interaction helps us remember how to put the term in a regression model. We form the interaction term as the product of the variables representing the main effects. In our example, miles and time of day represent the main effects.

In the example above, if we limit our model statement to the main effects, we can see that the situation would be described by parallel lines:

Model statement (with main effects only): $\text{Time} = \beta_0 + \beta_1\text{Miles} + \beta_2\text{TimeOfDay} + \varepsilon$.

If TimeOfDay is coded as 0 for Mid-day and 1 for Rush Hour, we can see the following:

(for travel mid-day): $Time = \beta_0 + \beta_1 Miles + \beta_2 TimeOfDay + \varepsilon$
$= \beta_0 + \beta_1 Miles + \beta_2(0) + \varepsilon$
$= \beta_0 + \beta_1 Miles + \varepsilon$ [a straight line with slope β_1]

(for rush hour travel): $Time = \beta_0 + \beta_1 Miles + \beta_2 TimeOfDay + \varepsilon$
$= \beta_0 + \beta_1 Miles + \beta_2(1) + \varepsilon$
$= (\beta_0 + \beta_2) + \beta_1 Miles + \varepsilon$ [a straight line with slope β_1]

Both equations plot as a straight line with the same slope (β_1).

Adding the interaction term (Miles)(TimeOfDay) or M*TOD, for short, gives a new model statement:

Model statement (adding interaction): $Time = \beta_0 + \beta_1 Miles + \beta_2 TimeOfDay + \beta_3 M*TOD + \varepsilon$.

Again if TimeOfDay is coded as 0 for Mid-day and 1 for Rush Hour, we can see the following:

(for travel mid-day): $Time = \beta_0 + \beta_1 Miles + \beta_2 TimeOfDay + \beta_3 M*TOD + \varepsilon$
$= \beta_0 + \beta_1 Miles + \beta_2(0) + \beta_3 M*(0) + \varepsilon$
$= \beta_0 + \beta_1 Miles + \varepsilon$ [a straight line with slope β_1]

(for rush hour travel): $Time = \beta_0 + \beta_1 Miles + \beta_2 TimeOfDay + \beta_3 M*TOD + \varepsilon$
$= \beta_0 + \beta_1 Miles + \beta_2(1) + \beta_3 Miles(1) + \varepsilon$
$= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Miles + \varepsilon$ [a straight line with slope $\beta_1 + \beta_3$]

Both equations plot as a straight line but with different slopes. If $\beta_3 > 0$, the slope of the line for rush hour travel will be steeper than the line for mid-day travel.

Obtaining the Excel Output

Once we recognize that we need to consider the possibility of interaction in a model, we can start to write the model for the scenario. If the interaction is between two quantitative variables or if the interaction is between a quantitative variable and a qualitative variable where there is only one Dummy Variable needed, writing the model statement is straightforward. Include the variables that represent the main effects in the model and then add an interaction term created as the product of the variables from the main effects. If there are other predictor variables in the theory, be sure to include these also.

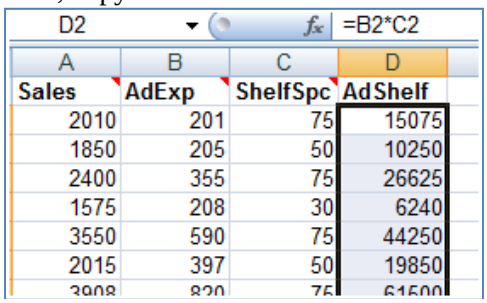
Once we have the model statement, we continue with the same steps as in the previous tutorials:

- Recognize the way Excel wants the data to be displayed in the with the Xs in consecutive columns in the spreadsheet
- Enter (or confirm) data in the needed format (This will include creating a column of data for the interaction variable)
- Use the Regression procedure in the Data Analysis Tools to obtain output
- Clean up the output
- Move on to the hard part…understanding what the output tells you

Example 1 (Interaction between two quantitative variables):
Case 1: Using the predicting sales example, consider the analysis that would be needed to address the following theory/question: Is the impact of advertising expenses the same regardless of the amount of visibility of the product in the store? Likewise, is the impact of additional shelf space the same regardless of the amount spent on advertising?

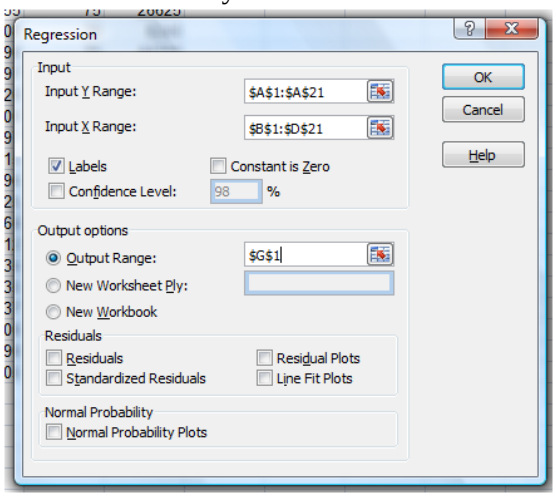
The theory talks about Advertising Expenses, Shelf Space, and the possibility that the impact of one of these variables on our prediction of sales may depend of the value of the other one of these variables. So we write the model statement as: $Sales = \beta_0 + \beta_1 AdExp + \beta_2 ShelfSpc + \beta_2 AdShelf + \varepsilon$ where AdShelf is the interaction between Advertising Expenses and Shelf Space.

We already have columns of data for Sales, AdExp, and ShelfSpc. We need a column for AdShelf next to the two other independent variables. Therefore, click in cell D1 and enter the variable name. To create the data, click in cell D2 to enter the first value. This is found by multiplying the value in B2 by the value in C2 (using the formula =B2*C2). Once this is done, copy the formula down the column.



	A	B	C	D
	Sales	AdExp	ShelfSpc	AdShelf
	2010	201	75	15075
	1850	205	50	10250
	2400	355	75	26625
	1575	208	30	6240
	3550	590	75	44250
	2015	397	50	19850
	3908	820	75	61500

Once this is done, you can access the Regression procedure in the Data Analysis Tools, select the data based on the information in your model statement and hit enter to obtain the output below.



SUMMARY OUTPUT

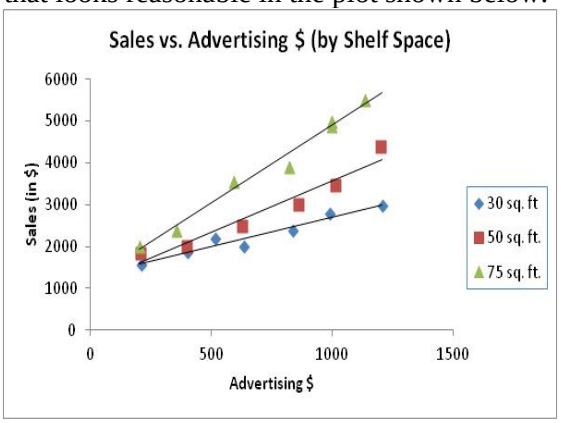
Regression Statistics	
Multiple R	0.989149
R Square	0.978416
Adjusted F	0.974368
Standard E	188.5514
Observatio	20

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regressior	3	25784705	8594902	241.7583	1.56E-13
Residual	16	568826.2	35551.64		
Total	19	26353531			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	1333.178	290.9994	4.581378	0.000307
AdExp	-0.15122	0.378646	-0.39938	0.694895
ShelfSpc	-2.62532	5.345963	-0.49109	0.630034
AdShelf	0.051954	0.006864	7.569264	1.13E-06

Be careful about interpreting the p values in the coefficient section. Recall that we added the interaction term <u>because</u> we believed that we could not interpret the coefficients of the main effects in isolation. The interaction term is allowing us to answer a question about moving from parallel lines to lines that are not parallel. A low p value for this term would imply that there is sufficient evidence to keep the term in the model and allow for lines that are not parallel—something that looks reasonable in the plot shown below.



Case 2: Using the final grade scenario, consider the theory: Does the impact (on the final grade) of missing class depend on how a student is performing in the class (as measured by the midterm grade)?

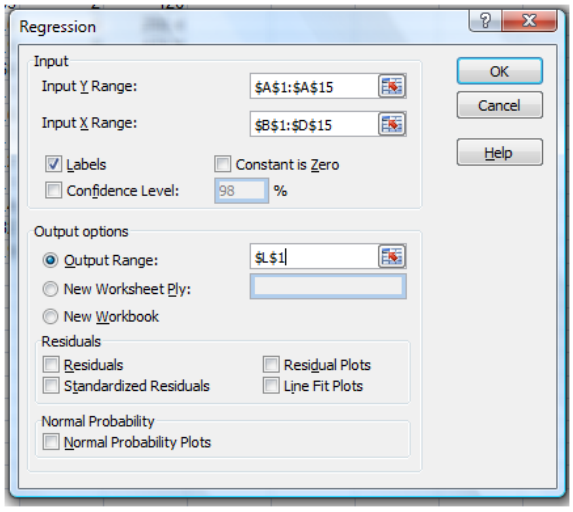
This theory is similar to the previous one. We will use a model statement that is set up the same way:

$$\text{Final} = \beta_0 + \beta_1\text{Midterm} + \beta_2\text{Absences} + \beta_2\text{MidAbs} + \varepsilon$$

where MidAbs is the interaction term between Midterm grades and Absences.

Like before, we will create a column with the interaction term, select the data in the Regression procedure, and obtain the output. Each step is shown below.

A	B	C	D
Grade	Midterm	Absences	MidAbs
57.25	32.2	2	64.4
92.65	84	1	84
71.25	63	2	126
70.95	72.6	4	290.4
81.95	85.9	2	171.8
81.5	68	2	136
93.7	93.1	1	93.1
87.5	75.6	2	151.2
92.45	95.1	1	95.1
80.55	76.2	1	76.2
89.55	92.1	1	92.1
77.85	72.4	1	72.4
72.55	82	4	328
91.1	78.9	0	0



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.940601
R Square	0.88473
Adjusted F	0.850149
Standard E	4.137316
Observatio	14

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regressior	3	1313.808	437.9361	25.58429	5.24E-05
Residual	10	171.1738	17.11738		
Total	13	1484.982			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	38.78853	25.22175	1.5379	0.155091
Midterm	0.640433	0.309672	2.0681	0.065492
Absences	2.378107	13.37134	0.177851	0.862391
MidAbs	-0.08159	0.168776	-0.4834	0.639221

This time the interaction term has a high p value—indicating that we do not have sufficient evidence to include the variable in the model. In order to write the regression equation, we would go back to the regression output that was generated to correspond to the model statement:

$$\text{Final} = \beta_0 + \beta_1\text{Midterm} + \beta_2\text{Absences} + \varepsilon$$

Example 2 (Interaction present between a quantitative and a qualitative variable):
Using the Weight Scenario, consider the following theory:

- Knowing someone's height and gender is useful; but to describe how a change in one characteristic relates to an expected change in weight depends on the other characteristic. Said another way, the expected difference in weights of two people of the same height but different gender would depend on the height. Or, to predict the expected weight gain for an individual as they get taller would depend on

whether you were predicting for a male or female. Or yet another way to say this: Males and females would not be expected to gain weight at the same rate as they get taller.

This time we have one quantitative independent variable (height) and one qualitative independent variable (gender). Since there are only two possible outcomes for gender, we only need one Dummy variable to represent gender. As in the previous tutorial, we will define GenderF = 1 for Females and 0 for Males. From here, we will proceed the same way that we did in Example 1 for this tutorial.

We will write the model statement as: Weight = $\beta_0 + \beta_1$Height + β_2GenderF + β_3HtGenderF + ε where GenderF is the interaction between Height and Gender. Then we will create the needed column of data, use the Regression procedure to obtain output, and look to see if the interaction term is needed. The data and the output are shown below.

=H2*I2

G	H	I	J
Weight	Height	GenderF	HtGenderF
140	65	1	65
125	64	1	64
115	63	1	63
190	73	0	0
125	64	1	64
250	76	0	0
125	61	1	61
185	66	0	0
200	71	0	0
155	73	0	0
100	65	1	65

SUMMARY OUTPUT

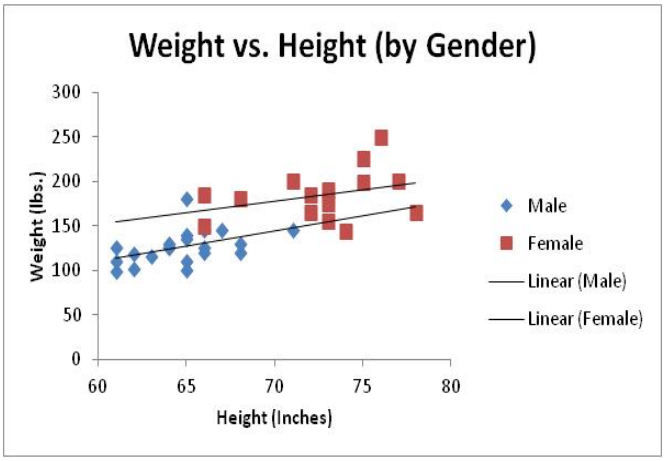
Regression Statistics	
Multiple R	0.823826
R Square	0.67869
Adjusted F	0.648567
Standard E	22.02669
Observatio	36

ANOVA

	df	*SS*	*MS*	*F*	*Sig.F*
Regressior	3	32794.03	10931.34	22.53072	4.98E-08
Residual	32	15525.6	485.1751		
Total	35	48319.64			

	Coefficients	*Standard Err*	*t Stat*	*P-value*
Intercept	-2.04023	117.5185	-0.01736	0.986256
Height	2.568966	1.616815	1.588905	0.121915
GenderF	-94.3619	170.8611	-0.55227	0.584599
HtGenderF	0.86624	2.506336	0.34562	0.73189

According to the p value associated with the HtGenderF row, we do not have sufficient evidence to include the interaction term in the model. If we were to plot the equations suggested by this output on a graph, we would see that the lines are almost parallel. Relative to the variation around the lines, the visual evidence is consistent with the regression output. Reverting back to the output from the model: Weight = $\beta_0 + \beta_1$Height + β_2GenderF + ε appears to be appropriate.



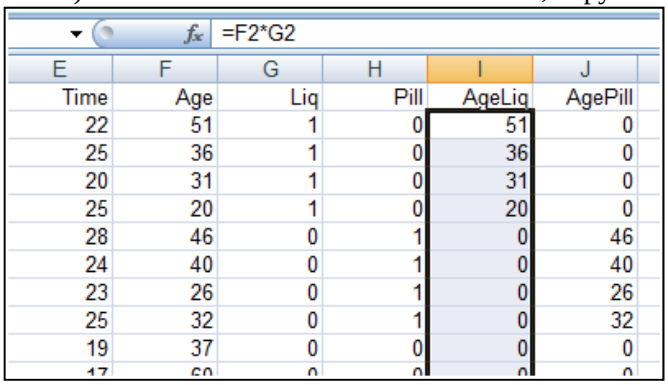
Example 3 (Interaction between a quantitative and qualitative variable where there are more than two possible outcomes for the qualitative variable):

Using the Time to Relief Scenario to test if "the choice of most effective method of administration depends on the age of the patient" presents a new challenge. In this case, the theory talks about a possible interaction between the Age of the individual and the Method of administration of the drug. The complicating factor is that there are three ways that the drug can be administered which leads to the need to define two Dummy variables to represent "Method." In the tutorial on working with qualitative variable, remember that we defined the following:

Pill = 1 if the drug is administered in Pill form and 0 otherwise
Liq = 1 if the drug is administered in Liquid form and 0 otherwise
We did not need a variable for Shot since Pill = 0 and Liq = 0 would tell us that the method was "shot."

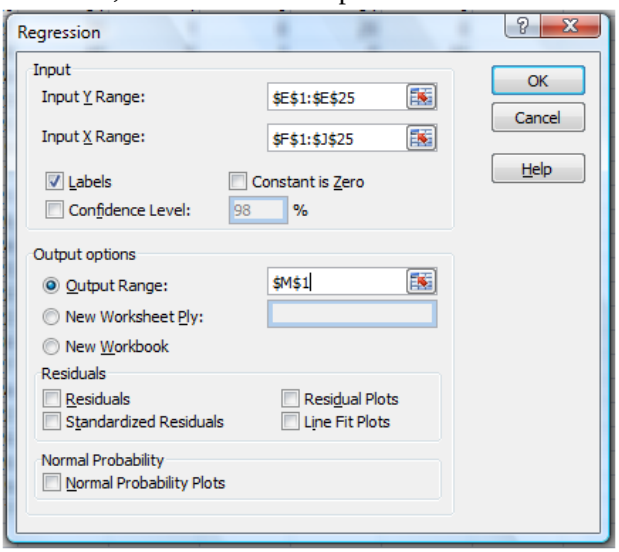
In our analysis, any question involving Method had to keep the Pill and Liq variables together. To create the Age by Method interaction, we will multiply Age by each of the Dummy variables to create two new variables. Like the two variables for "Method," the two new variables for the interaction between Age and Method will have to stay together—either both of these new variables will be in the model or both will be out of the model.

The model will be: $Time = \beta_0 + \beta_1 Age + \beta_2 Pill + \beta_3 Liq + \beta_3 AgePill + \beta_4 AgeLiq + \varepsilon$ where AgePill and AgeLiq together represent the interaction between Age and Method. From here, we create the two new columns. The first entry in the AgePill column is created by multiplying the corresponding value for Age by the value for Pill (=F2*G2). Likewise, the first entry in the AgeLiq column is created by multiplying the corresponding value for Age by the value for Liq (=F2*H2). Once the first row of data are created, copy the formulas down the column.



E	F	G	H	I	J
Time	Age	Liq	Pill	AgeLiq	AgePill
22	51	1	0	51	0
25	36	1	0	36	0
20	31	1	0	31	0
25	20	1	0	20	0
28	46	0	1	0	46
24	40	0	1	0	40
23	26	0	1	0	26
25	32	0	1	0	32
19	37	0	0	0	0

Next, access the Regression procedure from the Data Analysis Tools, select the data as indicated in the model statement, and obtain the output as shown below.



M	N	O	P	Q	R
SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.95213				
R Square	0.906552				
Adjusted R Squa	0.880594				
Standard Error	1.130002				
Observations	24				
ANOVA					
	df	*SS*	*MS*	*F*	*Sig. F*
Regression	5	222.9741	44.59481	34.92418	1.17E-08
Residual	18	22.98427	1.276904		
Total	23	245.9583			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	
Intercept	22.86044	1.226579	18.63755	3.24E-13	
Age	-0.09609	0.029831	-3.22106	0.004738	
Liq	2.471817	1.81216	1.364017	0.189377	
Pill	-5.2618	1.822075	-2.88781	0.0098	
AgeLiq	0.027354	0.046447	0.588932	0.563225	
AgePill	0.300059	0.045046	6.661215	2.99E-06	

This time we cannot simply read off the answer to whether interaction should be included in the model. The p values in the rows for AgeLiq and AgePill do not take into account that these two variable must be treated as a pair. Description of how to find the appropriate p value will be addressed in class.

If we simplify the regression equation suggested by the output to find a line for each method of administration of the drug as a function of the age of the person, we can plot these lines on a graph. From this, we see that the lines do tend to follow the dots that correspond to the different methods and, clearly, the lines are not parallel to each other. We should not be surprised when the calculations for a p value associated with the interaction terms is small—indicating a need to allow for lines that are not parallel!