Multiple Linear Regression Excel 2010Tutorial
For use when at least one independent variable is qualitative

This tutorial combines information on how to obtain regression output for Multiple Linear Regression from Excel (when qualitative variables are included) 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 Linear Regression (basic) tutorial.

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 Reg3 Excel file for this tutorial is located at http://faculty.ung.edu/kmelton/Data/Reg3.xlsx. The Excel file for this tutorial contains data on two sheets accessed at the bottom left of the page. Each sheet is related to one of the scenarios described in the Word document.

Obtaining Multiple Linear Regression Output (when a variable is qualitative – i.e., categorical)

In the previous tutorials, we started with the model statement. That is problematic here, since each variable in the model statement should have numerical values. This major difference between obtaining output for models with qualitative variables must be address at the very start of the analysis. How do you change a variable that is not quantitative into something that Excel can use to crunch the numbers? We need to use a method that will provide results that reflect the relationships that exist between the dependent variable and the "levels" of the qualitative variable. Also, the method must provide results that are repeatable—i.e., a different researching using the same data set (same raw data) should get the same results.

[Warning 1: The description in the paragraph does not lead to correct results.] Many people are tempted to define a single variable and assign numbers to each "level" of the qualitative variable. For example, if we were trying to include "Home Campus" as a predictor variable where students could be identified as being from any one of four home campuses, we might be tempted to say let Campus = 1 for location A; 2 for location B; 3 for location C; and 4 for location D. Following this logic, the model statement would include a variable called Campus and the values 1, 2, 3, and 4 would be entered into Excel. This sounds easy enough, but if two researchers code the campuses differently, the results may be totally different! We need to find a better way.

[Warning 2: The description in this paragraph also leads to problems.] Another suggestion would be to identify a new variable for each level of the qualitative variable and assign one of two values to each new variable (one value for true and another for false). Assigning values of 1 for true and 0 for false would be consistent with many other commands in Excel and in computer operations in general. Using our previous example, the categorical variable "Home Campus" would translate into four quantitative variables for analysis—LocA, LocB, LocC, and LocD. In this case if a student's home campus is Location A, the row of data for this student would show a value of 1 for LocA and a value of 0 for LocB, LocC, and LocD. When you ask Excel to crunch the numbers, you will see some strange output (and strange can be seen in a variety of ways). Sometimes you will see a period (.) or a zero (0) where you expect to see numbers; other times you will see #NUM! where you expect to see a p value. What has happened is that you have given Excel too much information!

[Finally, something that will work.] The previous approach was on the correct path. Rather than creating a new variable for each of the levels of the qualitative variable, we should drop one of these new variables. Therefore, if there are n levels of the qualitative variable, we should define n-1 new variables. These new variables are called "Indicator Variables" or "Dummy Variables," and are coded with 0 or 1. When we add the new variables to our model statement, we MUST remember that all of these new

variables together represent the one conceptual variable. Therefore, any analysis about whether the conceptual variable is needed must consider these variables as a group (rather than individually).

Think about our previous example, if you know the values for three of the location variables, isn't the value for the other one obvious? That fourth piece of information is redundant. Giving redundant information causes problems in the math (similar to trying to divide by zero). So we could add three variables LocA, LocB, and LocC (each with its own coefficient) to our data set to represent the four home campuses. Then any theory that relates to home campus with have all three of these variables in the model (if campus is significant) and would have none of these three variables in the model (if campus is not significant). A model with one or two of the Loc variables would not make sense (since this would be like having part of the conceptual variable in the model and part of it out).

Now the steps to obtain the output:

- Identify any qualitative variable(s) in your theory. For each qualitative variable, determine (n) the number of levels of the variable and define n-1 Dummy Variables to represent that qualitative variable where each Dummy Variable will be assigned a value of 0 or 1 in the spreadsheet. Naming the Dummy Variables so that you can remember what they represent is suggested. For example, saying "X_1 = 1 if the student's home campus is Location A and 0 if not" is more difficult to keep up with than saying "LocA = 1 if the student's home campus is Location A and 0 if not."
- Write your model statement including the Dummy Variables. Be sure to keep a "note to self" that all n-1 Dummy Variables used to describe a qualitative variable must be viewed together.
- From here, you will continue with the same steps as in the previous tutorial
    - 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
    - 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 (Where the independent variables include one quantitative and one qualitative predictor variable—and where there are only two outcomes possible for the qualitative variable): Using the predicting weight example ["Weight" worksheet accessed at the lower left], consider the analysis that would be needed to address the following theory/question: Should we consider both height and gender when predicting someone's weight?

To evaluate this theory, we need to recognize that gender is a qualitative variable with two possible outcomes (Male or Female). Since there are two possible outcomes, we need to define one "Dummy Variable." Suppose, we define GenderF to be 1 if we are talking about a Female and 0 if we are talking about a Male. This would lead to the following model statement to guide our analysis:

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

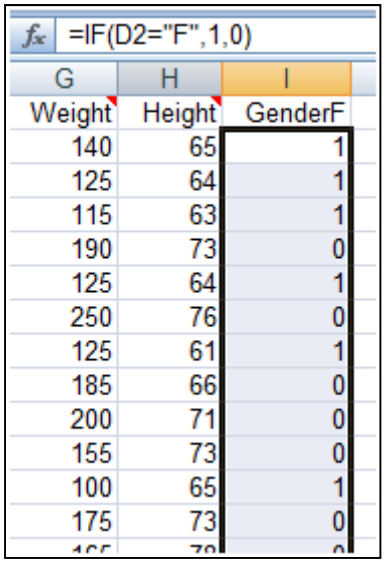
[After we get output for this model, we will see how the output would differ if we had defined our Dummy Variable in terms of Males.]

From the Model statement we recognize that we need a column for Weight, a column for Height, and a column for GenderF—and the columns for Height and GenderF need to be next to each other. You will see that the Weight and Height data have been copied into columns G and H; and the name for the new variable has been entered in cell I1. We could use a "brute force" approach to assign a 1 when the value column D is F; and assign a 0 when the value in column D is M. With small data sets this may be

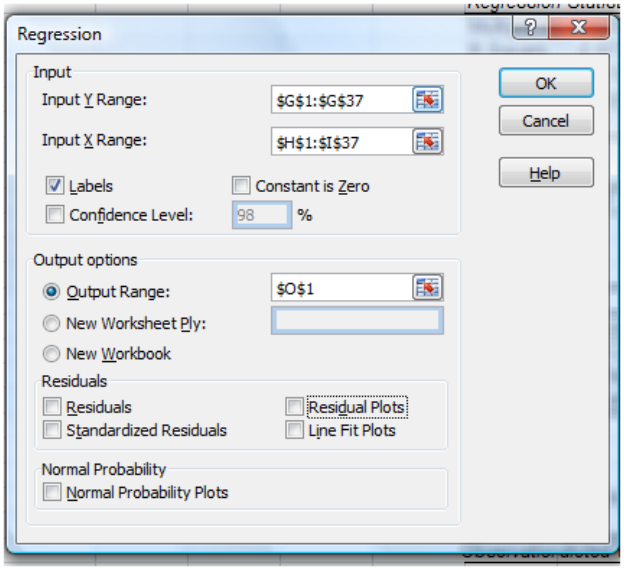
reasonable, but for large data sets this is time consuming and increases the likelihood of mistakes. Instead, we will let Excel do the work for us using the IF function. The function is set up to include three pieces of information inside the parentheses:

=IF(comparison, value if true, value if false)

In this example, we want to look at the value in column D to see if that value is the letter F. If the value is an F, we want to put a 1 in the same row of column I. If the value is not an F, we want to put a 0 in the same row of column I. Since we will be asking Excel to look for an F (characters) rather than numbers, we will need to put the F in quotes ("F"). To enter the function in the first cell for GenderF, click on cell I2 and enter =IF(D2="F",1,0) and hit enter. This tells Excel, "If cell D2 has the letter F in it, put a 1 in the current cell; otherwise put a 0 in the current cell. Then go back to cell I2 and copy/fill the function down the column to populate the data in the column. Once this is done, you are ready to select the data to obtain the regression output. The following two figures show the data and the Regression Dialogue box. Note that the model statement shows Weight as Y and Height and GenderF as X, so this identifies the data to select in the dialogue box.



G	H	I
Weight	Height	GenderF
140	65	1
125	64	1
115	63	1
190	73	0
125	64	1
250	76	0
125	61	1
185	66	0
200	71	0
155	73	0
100	65	1
175	73	0



The following output appears when you click OK.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.823098
R Square	0.67749
Adjusted R Squ	0.657944
Standard Error	21.73083
Observations	36

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	32736.08	16368.04	34.66123	7.78E-09
Residual	33	15583.56	472.2291		
Total	35	48319.64			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	-28.211	88.66431	-0.31818	0.752355
Height	2.929445	1.218824	2.403502	0.022017
GenderF	-35.4611	12.09736	-2.93131	0.006088

Remember that you only needed one variable to account for both genders (GenderF takes on a value of 1 if we are talking about females and a value of 0 if we are talking about males). Therefore, the one regression equation identified in the coefficient section allows estimates for males and females:

General equation: Weight(hat) = -28.211 + 2.929Height – 35.461GenderF

For Males GenderF=0 Weight(hat) = -28.211 + 2.929Height – 35.46(0)
= -28.211 + 2.929Height
For Females GenderF=1 Weight(hat) = -28.211 + 2.929Height – 35.461(1)
= -63.672 + 2.929Height

Using this model statement produces two lines (one for males and one for females) and the lines are parallel to each other (since they have the same slope but different intercepts). We will address the question about whether these lines should be parallel to each other in another tutorial.

Lingering question 1: Wasn't the choice of 1 for Female arbitrary? What if we had worked the problem the "other way around"? What if we had coded our gender variable as 1 for Males and 0 for Females? In the following output GenderM = 1 if the individual is a male and 0 if the individual is a female.

Weight = $\beta_0 + \beta_1$Height + β_2GenderM + ε

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.823098
R Square	0.67749
Adjusted R Squ	0.657944
Standard Error	21.73083
Observations	36

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	32736.08	16368.04	34.66123	7.78E-09
Residual	33	15583.56	472.2291		
Total	35	48319.64			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	-63.6721	79.01772	-0.8058	0.426129
Height	2.929445	1.218824	2.403502	0.022017
GenderM	35.46114	12.09736	2.931313	0.006088

If you compare this output to the output using GenderF, you will see that the Summary Output and the ANOVA table are identical; and part of the coefficient section "matches". But some of the entries in the coefficient section look a little different. The differences are not related to whether a variable is significant (or not) nor related to the estimates for weights of individuals with the same height and gender characteristics. Simply the equation in this output to show how each gender's weight would be estimated and you will find:

For Males Weight(hat) = -28.211 + 2.929Height and
For Females Weight(hat) = -63.672 + 2.929Height

The coefficient of -35.461 for GenderF in the first output says that for two individuals of the same height (but different gender), the female would be expected to weigh 35.461 pounds less than the male. The coefficient of 35.461 for GenderM in the second output says that for two individuals of the same (but different gender), the male would be expected to weigh 35.461 pounds more than the female. This is two ways of saying the same thing!

Lingering question 2: What would the output look like if I had put in a variable for each gender? The following output uses the model statement: Weight = $\beta_0 + \beta_1$Height + β_2GenderF + β_3GenderM + ε. Note the row for GenderM in the coefficient section—a clear indication "you goofed."

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.823098
R Square	0.67749
Adjusted R Squ	0.627641
Standard Error	21.73083
Observations	36

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	3	32736.08	10912.03	34.66123	3.6E-10
Residual	33	15583.56	472.2291		
Total	36	48319.64			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	-28.211	88.66431	-0.31818	0.752355
Height	2.929445	1.218824	2.403502	0.022017
GenderM	0	0	65535	#NUM!
GenderF	-35.4611	12.09736	-2.93131	0.006088

Other Examples:

Example 2 (Where the only independent variable is a qualitative variable with more than two possible outcomes for that variable) Predicting time to relief for a medicine ["Time" Worksheet accessed at the lower left of the same file.]: In this example, we want to see if the method of administration of a drug can be used to estimate the time that it will take for the patient to feel relief. The drug can be administered in three different forms: a Pill to be swallowed, a Liquid to be swallowed, or as a shot. All times are measured from the time that the patient received the medicine until that same patient reaches some specified level of better (e.g., temperature drops to normal).

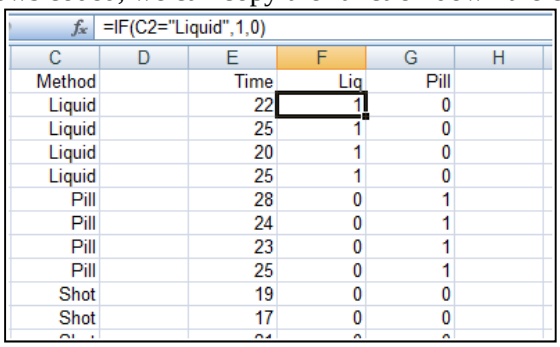
Since method of administration of the drug is a qualitative variable and there are there are three ways the drug can be administered, we will need two Dummy Variables. If we define the following:

Pill = 1 if the drug is administered in pill form; and Pill = 0 if not in pill form

Liq = 1 if the drug is administered in liquid form; and Liq = 0 if not in liquid form

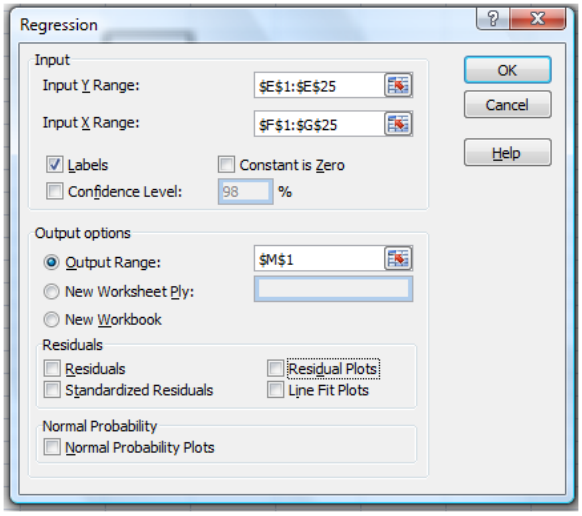
Then we can use the following model statement: Time = $\beta_0 + \beta_1 Pill + \beta_2 Liq + \varepsilon$

Based on this model statement, we will need a column for Time, a column for Liq, and a column for Pill. Remember that the data were originally provided in a table. The current file has each variable in a separate column and the data for Time have been copied into column E and the headings of Liq and Pill have been added in columns F and G. To assign the 0s and 1s to these two columns, we will use the IF function. In cell F2, we will look to see if the value in cell C2 is "Liquid. If so, Excel should enter a 1; otherwise Excel should enter a 0. Likewise in cell G2, we will look to see if the value in cell C2 is "Pill." Once we have the first data rows coded, we can copy the function down the columns.



C	D	E	F	G	H
Method		Time	Liq	Pill	
Liquid		22	1	0	
Liquid		25	1	0	
Liquid		20	1	0	
Liquid		25	1	0	
Pill		28	0	1	
Pill		24	0	1	
Pill		23	0	1	
Pill		25	0	1	
Shot		19	0	0	
Shot		17	0	0	

From here, the possess to obtain output is the same as in the previous regression examples. The model statement shows the Y variable isTime and the X variables are Liq and Pill. Selecting these in the dialogue box and clicking OK provides the following.



SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.802334				
R Square	0.64374				
Adjusted R Squ	0.609811				
Standard Error	2.042699				
Observations	24				
ANOVA					
	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	158.3333	79.16667	18.9729	1.97E-05
Residual	21	87.625	4.172619		
Total	23	245.9583			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	
Intercept	19.125	0.722203	26.48147	1.28E-17	
Liq	3.75	1.021349	3.671613	0.001421	
Pill	6.25	1.021349	6.119355	4.51E-06	

This time we will need to be careful about interpreting the p values in the coefficient section. The variables Liq and Pill <u>together</u> represent the three methods of administration of the drug. Remember if Liq = 0 and Pill = 0, the drug was administered in Shot form.

<u>But what if</u>? In the previous output we omitted Shot when we defined our Dummy variables. What would have happened if you had "omitted" a different method of administration? Consider the output from two other models that could have been used:

Time = $\beta_0 + \beta_1$Pill + β_2Shot + ε [omits Liquid] output on the left below

Time = $\beta_0 + \beta_1$Shot + β_2Liq + ε [omits Pill] output on the right below

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.802334				
R Square	0.64374				
Adjusted R Squ	0.609811				
Standard Error	2.042699				
Observations	24				
ANOVA					
	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	158.3333	79.16667	18.9729	1.97E-05
Residual	21	87.625	4.172619		
Total	23	245.9583			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	
Intercept	22.875	0.722203	31.67391	3.26E-19	
Pill	2.5	1.021349	2.447742	0.023252	
Shot	-3.75	1.021349	-3.67161	0.001421	

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.802334				
R Square	0.64374				
Adjusted R Squ	0.609811				
Standard Error	2.042699				
Observations	24				
ANOVA					
	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	158.3333	79.16667	18.9729	1.97E-05
Residual	21	87.625	4.172619		
Total	23	245.9583			
	Coefficients	*tandard Err*	*t Stat*	*P-value*	
Intercept	25.375	0.722203	35.13554	3.84E-20	
Shot	-6.25	1.021349	-6.11935	4.51E-06	
Liq	-2.5	1.021349	-2.44774	0.023252	

Any one of the three outputs will provide the same results. All three of the outputs give the same Summary Output, the same ANOVA output, and the same estimates for time to relief (with a little simplification of the equation).

<u>Signals that "you goofed."</u> If you were to include a Dummy Variable for all three methods of administration, you would receive the same #NUM! in one of the rows that we saw in the previous example.

A less obvious (but more serious) way to "goof" is when your output will not give you any clear signal that you have made an error. Look at the output from running the regression procedure twice using different codes for Method. Both use the exact same "raw" data. For the output on the left, method of administration of the drug was coded 1 for pill, 2 for liquid, and 3 for shot. For the output on the right, method of administration of the drug was coded 2 for pill, 1 for liquid, and 3 for shot. Excel had no way to know Method represented categorical data (Nominal scale data); so Excel "crunched" the numbers as though they represented measurements (Interval or Ratio scale data).

SUMMARY OUTPUT

NOTE: This is a WRONG way to analysis this problem. Method is a qualitative variable with 3 levels.

Regression Statistics	
Multiple R	0.797038
R Square	0.63527
Adjusted R Squ	0.618692
Standard Error	2.01932
Observations	24

Coding
Pill = 1
Liquid = 2
Shot = 3

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	1	156.25	156.25	38.31863	3.13E-06
Residual	22	89.70833	4.077652		
Total	23	245.9583			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	28.70833	1.090557	26.32446	3.99E-18
Method	-3.125	0.50483	-6.1902	3.13E-06

SUMMARY OUTPUT

NOTE: This is a WRONG way to analysis this problem. Method is a qualitative variable with 3 levels.

Regression Statistics	
Multiple R	0.318815
R Square	0.101643
Adjusted R Squ	0.060809
Standard Error	3.169158
Observations	24

Coding
Pill = 2
Liquid = 1
Shot = 3

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	1	25	25	2.489157	0.128905
Residual	22	220.9583	10.04356		
Total	23	245.9583			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	19.95833	1.711541	11.66103	6.88E-11
Method	1.25	0.792289	1.577706	0.128905

Based on the output on the left (and the p value of 3.13×10^{-6}), we would conclude that <u>we have</u> sufficient evidence to say that the method of administration of the drug can be used to predict time to relief. Based on the output on the right (and the p value of 1.28905), we would conclude that <u>we do not have</u> sufficient evidence to say that the method of administration of the drug can be used to predict time to relief.

Also as seen in the table that follows, if you used the regression equation from the output for each approach to coding, you would obtain different estimates for the expected time needed for the drug to be effective. And you would reach a different conclusion about which method of administration would be the fastest to provide relief!

	Output on left	**Output on right**
Regression equation	Time(hat) = 28.70833 – 3.125Method	Time(hat) = 19.95833 + 1.25Method
Estimate for pill	Time(hat) = 28.70833 -3.125(1) = 25.58333	Time(hat) = 19.95833 + 1.25(2) = 22.45833
Estimate for liquid	Time(hat) = 28.70833 -3.125(2) = 22.48533	Time(hat) = 19.95833 + 1.25(1) = 21.20833
Estimate for shot	Time(hat) = 28.70833 -3.125(3) = 19.33333	Time(hat) = 19.95833 + 1.25(3) = 23.70833

The same raw data should produce the same conclusions—in terms of whether a variable is significant, in terms of predicted outcomes, and in terms of relative rankings of the levels of the qualitative variable! The outcome obtained should not be determined by the arbitrary choice of values assigned to a qualitative variable.

<u>Example 3</u> (Where the independent variables include one quantitative variable and one qualitative variable with more than two possible outcomes) An extension of the previous situation where we were predicting Time to relief: In this case, someone has suggested that we should consider age and the method of administration of the drug when predicting time to relief.

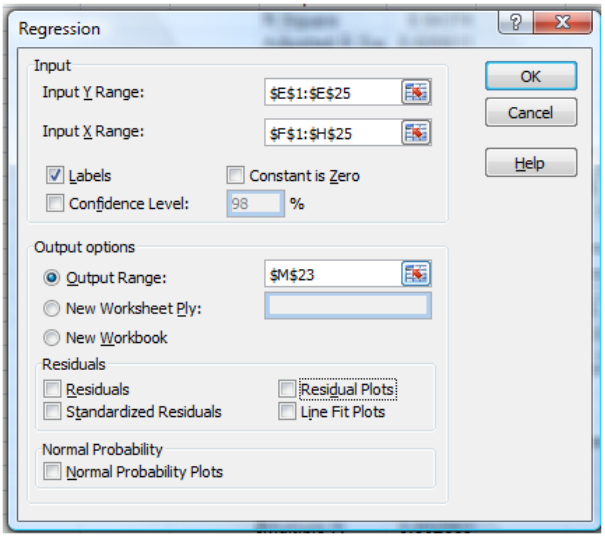
Like before, we will need to recognize that method of administration of the drug is qualitative and that we need two Dummy Variables to represent the three methods of administration. This time we need to add age to the model statement. Therefore, we can write the model statement as:

$$\text{Time} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Pill} + \beta_3\text{Liq} + \varepsilon$$

Note: The order that the terms appear in the model statement is not important. Also, the terms β_0, β_1, β_2, … are representative of the value of the coefficient in the given model (and should not be assumed as "tied" to a specific variable or value across different models).

In order to obtain regression output, we will need a column of data for Time, one for Age, one for Pill and one for Liq. The illustration below shows Liq and Pill moved to columns G and H and the data for Age copied into column F so that the X variables are listed in the same order as the model statement. If you keep Liq and Pill in columns F and G and copy Age into column H, you will obtain the same results as the output in the illustrations here—the only difference will be the order of the rows in the coefficient section.

E	F	G	H
Time	Age	Liq	Pill
22	51	1	0
25	36	1	0
20	31	1	0
25	20	1	0
28	46	0	1
24	40	0	1
23	26	0	1
25	32	0	1
19	37	0	0
17	60	0	0
21	25	0	0
20	38	0	0
24	20	1	0
23	35	1	0



Clicking OK will produce the following output. Be careful when interpreting the results. Age does represent a quantitati ve variable, but you must remember that Liq and Pill <u>together</u> represent the qualitative variable Method of Administration.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.802663
R Square	0.644268
Adjusted R Squ	0.590908
Standard Error	2.091594
Observations	24

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	3	158.4631	52.82102	12.07403	9.87E-05
Residual	20	87.49527	4.374764		
Total	23	245.9583			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	18.89044	1.549934	12.1879	1.03E-10
Age	0.006034	0.035039	0.172202	0.865009
Liq	3.768856	1.051513	3.58422	0.001855
Pill	6.254525	1.046127	5.978744	7.59E-06

Signal that you "goofed": If you see the #NUM! in the coefficient section…you goofed. Also, if your qualitative variable has more than two levels (i.e., at least three different outcomes), using a single quantitative variable is a serious mistake that will not provide signals on the output that you goofed

Example 4 (Where the independent variables include quantitative and qualitative variables in some combination): Using the predicting weights example ["Weight" Worksheet accessed at the lower left of the same file.]

In this example we will consider the theory, "If we already know a person's height and gender, we should not need to consider their major when estimating weight. In this case we have two qualitative variables—gender and major. You may say, "Why do I need to worry about major since the theory says that we do not need to consider major?" Remember, theories are people ideas about what they believe—not all theories are correct. To see if knowing someone's major is not important, we need to consider the possibility that major is important. Therefore, we put major into the model …and hope that we do not find sufficient evidence to keep it there!

We have already created a Dummy Variable for gender. Since there were only two levels (Male and Female), we created one Dummy Variable. We will continue to use the GenderF that we created before.

We need to create Dummy Variables to represent a student's major. To do this we need to know what the possible outcomes were (or actually, we need to know how many different levels of the variable are represented in our data set). When the data were collected students responded to the question about major with five possible answers (Accounting, Finance, Management, Marketing, or Other). The data set does include some students from each group. Since there are five levels for the "Major" qualitative variable, we will need four Dummy Variables. We can define the following four Dummy Variables:

Acct = 1 for Accounting majors and 0 for all others
Finc = 1 for Finance majors and 0 for all others
Mgmt = 1 for Management majors and 0 for all others
Mktg = 1 for Marketing majors and 0 for all others

Therefore, if a student has a 0 in each of these four columns, they responded as "Other."

With this, we have enough to write our model statement to include height, gender, and major as predictors of weight: $Weight = \beta_0 + \beta_1 Height + \beta_2 GenderF + \beta_3 Acct + \beta_4 Finc + \beta_5 Mgmt + \beta_6 Mktg + \varepsilon$

We need to be careful when we use the IF statement to populate columns J, K, L, and M since the majors were entered in column E as Acc, Fin, Man, Mark, and Other. As we create the first data entry in each of these columns, our comparison statement will need to use the variable names as coded in column E. For example, the function in cell J2 will be =IF(E2="Acc",1,0) and cell K2 will be =IF(E2="Fin",1,0).

E
Major
Other
Other
Man
Man
Man
Man
Other
Other
Fin
Man
Acc
Other
Man
Fin
Mark

=IF(E2="Acc",1,0)

G	H	I	J	K	L	M
Weight	Height	GenderF	Acct	Finc	Mgmt	Mktg
140	65	1	0	0	0	0
125	64	1	0	0	0	0
115	63	1	0	0	1	0
190	73	0	0	0	1	0
125	64	1	0	0	1	0
250	76	0	0	0	1	0
125	61	1	0	0	0	0
185	66	0	0	0	0	0
200	71	0	0	1	0	0
155	73	0	0	0	1	0
100	65	1	1	0	0	0
175	73	0	0	0	0	0

Once the data are entered, obtaining the output becomes routine. Look at the model statement to see that Y is Weight and that the Xs include Height, GenderF, Acct, Finc, Mgmt, and Mktg. Select the data in the regression dialogue box to match this, and click OK to get the following output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83709
R Square	0.70072
Adjusted R Squ	0.638799
Standard Error	22.33068
Observations	36

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	6	33858.52	5643.086	11.31651	1.68E-06
Residual	29	14461.12	498.6594		
Total	35	48319.64			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	-31.1666	99.23415	-0.31407	0.755715
Height	3.110444	1.381684	2.251198	0.032122
GenderF	-41.7464	13.87516	-3.00872	0.005381
Acct	-2.4581	11.77758	-0.20871	0.836133
Finc	-20.201	14.80387	-1.36458	0.182882
Mgmt	-7.09388	10.73486	-0.66083	0.51394
Mktg	-13.7122	13.62509	-1.00639	0.32255

Again, be careful. All four of the variables related to major must be considered <u>together</u> since they are part of the same qualitative variable that has five possible outcomes. You cannot use the p values in the rows for these four Dummy Variables to draw conclusions about if major should be in the model. Interpretation of this output will be addressed in class.