*Module 3, Week 2, gretl problem set 6*

Let’s continue to explore our Boston housing dataset. Since there are 374 records in the processed dataset we don’t have to worry about the sample size. However, we do not know the population standard deviation. By now you should immediately know what that means!

Last week, Week 5, we considered a simple linear relationship between the average number of rooms in a home and the home’s value. This week we’ll start considering a multiple linear regression, i.e. we’ll add more independent variables to our model.

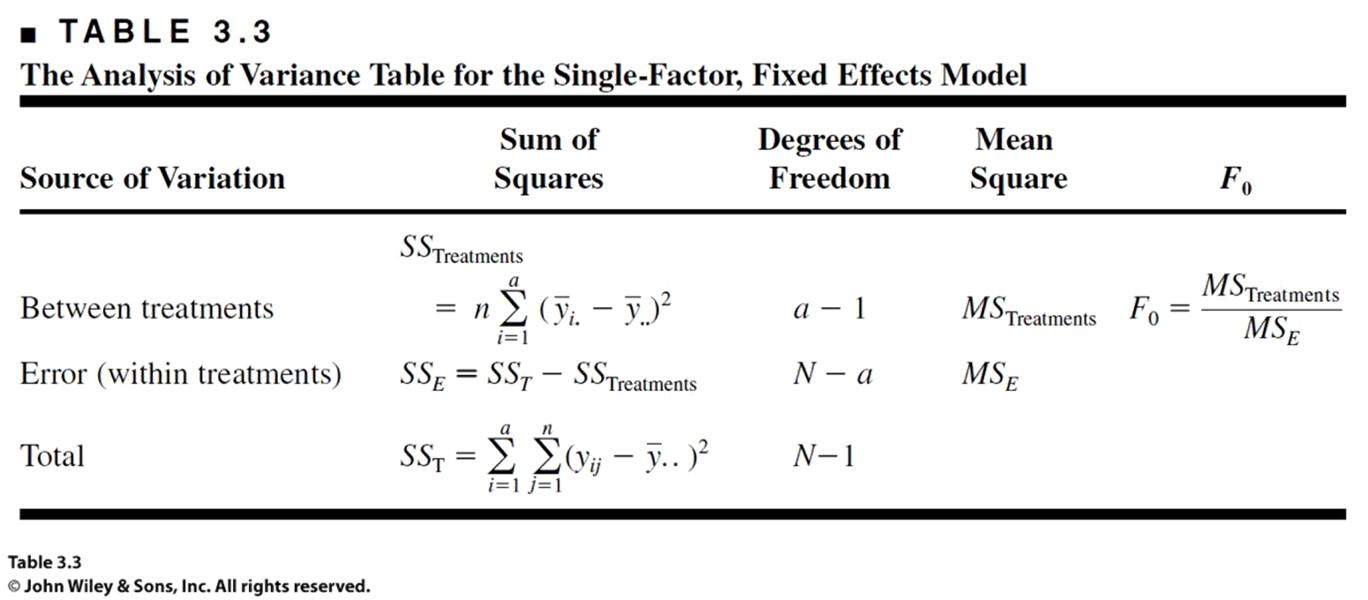
I’m going to skip any categorical variables. The reason for that will become more obvious when we start looking at ANOVA and logistic regression in detail. For more information about this you could look at Minitab’s blog at <https://blog.minitab.com/en/michelle-paret/regression-versus-anova-which-tool-to-use-when>. There may also be some confusion between simple (linear) regression, multiple variable regression and multivariate regression. To keep it simple just remember that:

**Simple Linear Regression** has one “Y” that must be continuous and one “X” variable:

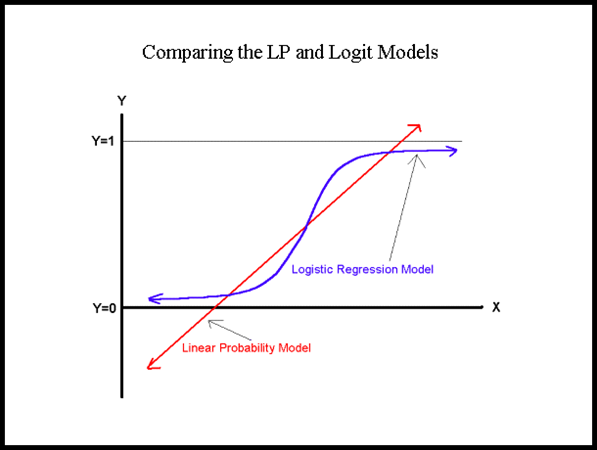
**Multivariable or multiple variable regression** has one “Y” that must be continuous and multiple “X” variables that may include categorical variables if they are encoded numerically:

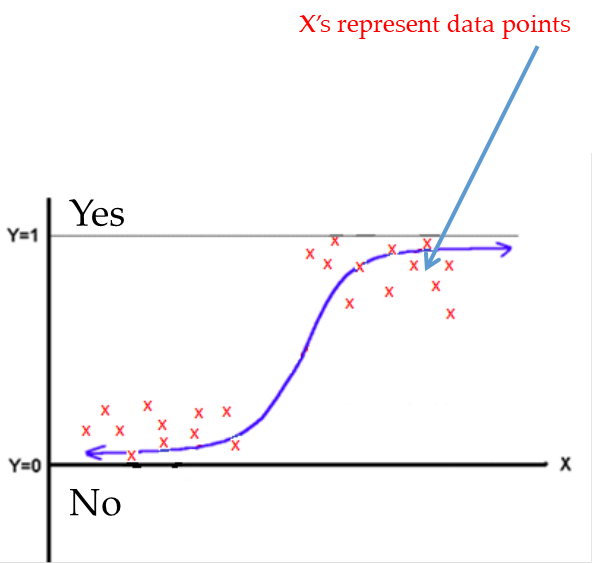
**Multivariate regression** has multiple “Y” variables which still much be continuous and multiple “X” variables that may include categorical variables if they are numerically encoded:

**ANOVA** can be used a lot but is more typically used with design of experiments to explore both within and across “group” or “treatment” variances. Here is a table for ANOVA for a single factor (single “X” or simple linear regression for us), fixed effects model (e.g. comparison across groups or treatments similar to or analogous to when we compare the means of two groups):



For now, consider **logistic regression**, which has a dependent “Y” categorical variable, with a simple graphic comparing a linear regression model and a logistic regression model. Logistic regression is often used as a basis in classification methods, e.g. does someone have cancer or not, yes or no.





Now that we’ve summarized all these different things we’ve been looking at, let’s get back to the problem we’re working on. Let's look at the Boston housing variables to see which we might include in a multiple variable or multivariable regression which are those that are most directly related to home values. I think that crime, whether or not there is a view of the Charles River, i.e. CHAS, nox, RM, AGE, DIS the distance to Boston's 5 primary employment centers, RAD the accessibility to radial highways, tax, ptratio the pupil-teacher ratio in local K-12 schools, and lstat the percentage of lower status of the population in the town are probably the most directly relevant.

Remember that the original authors were trying to prove that people were willing to buy homes in the Boston area, coming to somewhere that wasn't home and to pay more money for them because of the perception of less air pollution. They failed to prove that!

Just to recover the data from last week you can copy and paste the following commands to get the summary statistics and the ordinary least squares (OLS) model we built in gretl.

#Compute the descriptive stats for just the variables of interest

**summary** CMEDV RM

#Compute the correlation coefficient for the variables of interest

**corr** CMEDV RM

#Develop an ordinary least squares model for this data

**ols** CMEDV 0 RM **--vcv**

Now what do we need to do to include additional independent variables in our OLD model? In general it is as simple as listing the variables with the dependent variable first and including 0 to get the intercept.

#Develop a multiple variable OLS model for this data

**ols** CMEDV 0 crim nox RM AGE DIS TAX PTRATIO black lstat **--vcv**

Now, the output from gretl will show two models called Model 1 and Model 2. You can see that the output between these varies even though the same data was used. This is because by conducting a multiple variable or multivariable regression you have changed the method sufficiently that you are no longer looking at results for the same thing. That is, this is no longer the case where you have the same data and use two different methods to ensure that you are getting a single, correct response. Now, you have the same data but are looking for two different results, even though both methods involve regression.

Think about it this way. For the first case, the simple linear regression, there was only one variable that contributed to the slope of the regression line. Therefore, the output for Model 1 only lists one coefficient related to the slope based on the RM variable. For Model 2, there are many variables that contribute to the slope of the regression line. You can see how much each contributes by looking at the respective coefficients. By comparison, for Model 1 (the simple linear regression) the slope coefficient for RM is 7.99551 or about 8.00. For Model 2 (the multivariable regression) the slope coefficient for RM is 10.2867 or about 10.29. And, in both cases it is statistically significant. The R-squared value for Model 1 is 0.59 whereas for Model 2 it is 0.83. It seems that Model 2 explains a lot more of the variance in the data. But, not all factors, or independent variables, are statistically significant.

I’m uploading the gretlUser’s Guide for you in case you didn’t download that when you installed gretl. The explanation about conducting a regression analysis in gretl begins with Chapter 2. Beyond that it is a bit difficult to find specific details. For example, the asterisks at the end of lines for the different independent variables indicates the level of significance for that variable. There is actually a pseudo-standard associated with this. One asterisk indicated significance at the 10% level or . Two asterisks indicates significance at the 5% level or . Three asterisks indicates significance at the 1% level or . You can see a reference to this in the gretl User’s Guide on page 423 but that isn’t necessarily obvious…

1. Which independent variables are NOT statistically significant in Model 2? Select all that apply.
   1. crim
   2. nox
   3. RM
   4. AGE
   5. DIS
   6. TAX
   7. PTRATIO
   8. black
   9. lstat
2. Most of the independent variables are significant at the 1% level, i.e. . True or False?
   1. True
   2. False
3. Having a confidence level of 0.1 means that when sampling there is more likely that the mean will fall within the confidence interval than if the confidence level were 0.01. (Hint, if you are unsure consider the solution to Example 8.4 in your OpenStax Introductory Statistics textbook.) True or False?
   1. True
   2. False
4. The standard error of regression, which is different than the standard error of the mean, represents the average distance that observed values are from the regression line of the model. The standard error of regression in Model 2 is preferred to the standard error of regression in Model 1 because it reflects more precision and/or less distance on average between the computed regression line that represents the model and the actual, observed values of the data. True or False?
   1. True
   2. False

One of the most common methods to test whether or not the best or correct independent variables have been included in a regression model is basically a trial and error approach. That is, you start eliminating variables to see what difference each elimination makes in the results. It isn’t the most elegant approach but it is pretty fast and easy to understand. Let’s give it a try!

1. First, eliminate the independent variable “nox” from the model and enter the resulting R-squared value. With “nox” removed R-squared equals \_\_\_\_\_\_\_\_\_.
2. Next, replace “nox” in the model and remove the independent variable “lstat”. With “lstat” removed R-squared equals \_\_\_\_\_\_\_\_\_.
3. Considering our original multivariable model, Model 2, let’s remove several independent variables and see if that makes a bigger difference. This time remove “nox,” “lstat,” “black,” and “crim”. With those independent variables removed the R-squared value equals \_\_\_\_\_\_\_\_\_.

It looks like just removing one or even a few variables at a time doesn’t make much change in the R-squared value. In case you wondered how I picked the order of the variables to remove I considered their respective P-values. I started with “nox” because it was not statistically significant so should not change the results of the model. Let’s go the other way.

1. With the independent variable “RM” removed the R-squared value of the model (in this case Model 6) is \_\_\_\_\_\_\_\_\_.
2. The new model built when RM was removed has changed the significance of the other independent variables, e.g. lstat, which now has a P-value of (select the best answer below):
   1. 0.00
   2. 1.95e-43 or almost zero
   3. 1.95
   4. None of the above
3. Now, if we also remove the independent variable “lstat” the R-squared value of the model is \_\_\_\_\_\_\_\_\_.

Removing two independent variables, “RM” and “lstat,” causes our model to go from an R-squared value in the 0.8’s to a value in the 0.2’s. This is a big change. In fact, changing which variables are included, or not, can change a model by a little or a lot. You must be careful when you consider the variables to include in a model!

It is worth noting that when we included all the independent variables we thought directly affected home values that RM had the lowest P-value and nox as well as lstat had the highest P-values. That is, the impression is that RM is most significant and really affects the regression model so should make a big difference in our results. But, when we only removed only nox or lstat from the model it made very little difference. However, this is where the notion that the P-value somehow tells the strength of a variable with regard to hypothesis testing, etc. and we can use that as a indicator for a variable whether or not that variable is statistically significant. However, as I have said many times, “If something is not statistically significant then it is NOT.” It is simply due to random chance. If you doubt that consider the difference between Model 2 and Model 3 when we eliminated “nox” which is not statistically significant. The difference in R-squared is out to the fifth decimal place which when considering significant digits (if you don’t recall what those are look back through the announcements) probably isn’t a real difference any more.

1. Just as the R-squared value indicated more or less precision in our regression model, the standard error of regression or S.E. of regression values indicated consistently, respective more or less distance between the regression line of the model to actual, observed data values. True or False?
   1. True
   2. False
2. If we include the independent variables “RM,” “AGE,” “TAX,” and “PTRATIO” in our model we will still have an R-squared value (slightly) greater than 0.80. True or False?
   1. True
   2. False
3. Using the slope coefficients of the model when “RM,” “AGE,” “TAX,” and “PTRATIO” are included in our model then enter the values for these slope coefficients in the appropriate places in the following equation.
4. The value of the intercept, i.e. -41.56, means that the regression line has a steep, negative slope. True or False?
   1. True
   2. False