*Module 3, Week 1, Paper and Pencil Assignment 5*

Although this is a relatively long document, there is really not that much for you to do. However, this is important information and topics that you really need to spend some time thinking about and getting a good intuitive understanding of.

You should have completed your gretl assignment for Week 5 by now. If not you should do that gretl assignment first. That gretl assignment will generate a dateset of small sample size (30 observations) for you to use to complete this paper and pencil assignment. NOTE!!! You can use any tool to complete this paper and pencil assignment, e.g. Excel. You do not have to try to add, subtract, etc. data from 30 observations by hand. The point of this assignment is to make sure that you understand what you need to do to conduct this type of analysis. Also, if you are going to use something like Excel you’ll need to make sure you have the “Data Analysis” add-on installed.

1. Generate some descriptive statistics, the correlation between home values and average number of rooms in the reduced sample data, and a regression model for your reduced sample size data file, i.e. the 30 observations you sampled from the process Boston housing dataset. You can use Excel or gretl or any appropriate method to do this. Compare your results from this regression model to the one from the gretl assignment using the entire data sample. For this question there are no incorrect answers. As you should understand from our prior course work, the values you obtain will change with every different sample you obtain from the larger dataset.

First consider the descriptive statistics. I’m putting my answers here. Each of your answers will and should vary from mine!

* 1. Was the mean home value greater than, equal to, or less than the computed home value using the larger dataset? Enter greater than, equal to, or less than.
  2. Was the mean (average) number of rooms greater than, equal to, or less than the computed average number of rooms using the larger dataset? Enter greater than, equal to, or less than…

Now, let’s dig a little deeper into the analysis.

* 1. Was the correlation between home values and the average number of rooms greater than, equal to, or less than the correlation value you computed using the larger dataset? Enter greater than, equal to or less than.

So far, for me the analysis and trends look similar between the smaller sample and the larger dataset…

* 1. Was the intercept value with this smaller sample greater than, equal to, or less than the intercept value you obtained in your gretl assignment using the larger dataset? Enter greater than, equal to, or less than.
  2. Was the coefficient of the slope greater than, equal to, or less than the value obtained from the larger dataset? Enter greater than, equal to, or less than.

**So, which answers are correct!**

* 1. We are starting to get some interesting differences. Both the intercept and coefficient of the slope indicate that they are statistically significant using the smaller sample size but the actual values are different and going in different directions.

To look at this further you might want to go back to gretl and modify the number of samples you take from the larger dataset. For example, you might take 60, 90, 120, 150, etc. until you get back to the larger sample size of 374. Repeat looking at these values for each sample size and see how they vary and hopefully converge to the original solution you got on your gretl assignment! This affect has actually been studied in detail and has associated statistics and statistical principles. You should also understand why there are no right or wrong answers to these questions because they will all differ depending on which 30 random samples were obtained from the larger dataset.

Another point is that you can see how it is getting increasingly difficult to do these computations by paper and pencil. This is why it is important to learn how to use codes to do the “heavy lifting” for you. However, you must know what you need to do in order to use the codes correctly, hence the paper and pencil exercises! Answer yes here if you have read this text.

1. Interpret the estimated slope coefficient you computed from your reduced sample data file. Enter the home value you compute if you have 10 rooms. Note that everyone will get a different answer here too! I got 58.31 as my (estimated) home value which is more than double the mean value of CMEDV.
2. This is a question that should have gone immediately after generating the OLS regression model. The question is, “**Are all the requirements and assumptions for conducting an OLS analysis satisfied?**” After you have read through the text for this question and done the appropriate tests enter yes or no. Again, there is no correct or incorrect answer for this question.

How do you check this? There are several basic assumptions including; independence or multicollinearity, linearity, normality, and homoscedasticity. **Memorize these!**  You should always immediately check your data and a model to make sure all requirements and assumptions have been met!

The easiest way right now to check for **independence,** or multicollinearity, is by considering correlations. As we go further we will begin performing a Durbin-Watson test. With regard to correlations, there will obviously be a relationship between the response or dependent variable and the independent variables or else we don’t have anything to begin with. So, we can just leave the dependent variable out of computing the correlations. If two independent variables are highly correlated you will need to consider which to eliminate from the data! Here are the correlation values for my 30-sample dataset:

**Correlations between independent variables in the small Boston housing dataset**

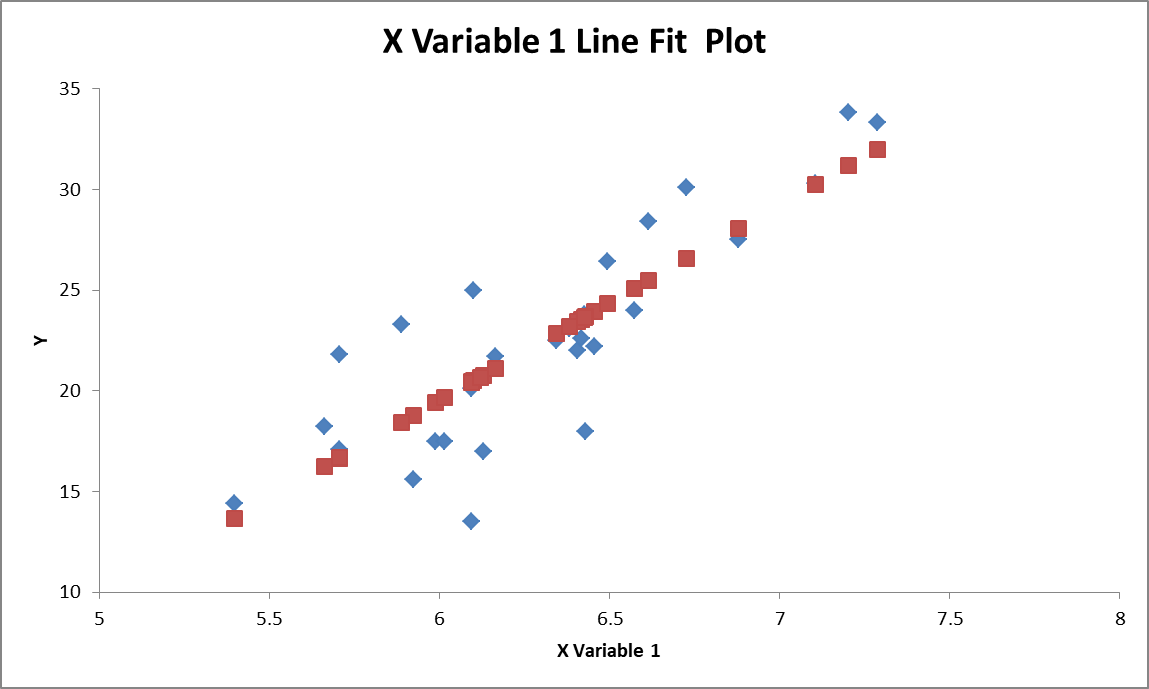


**Excel output for the OLS model using the small Boston housing dataset**

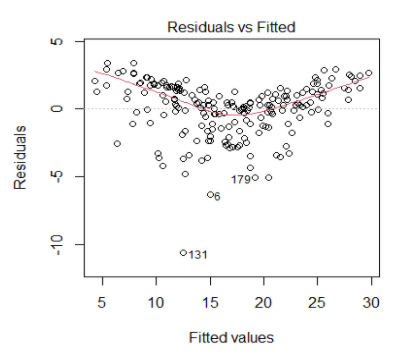


It looks like the highest correlations are between DIS and ZN, nox and INDUS, nox and AGE, lstat and AGE, crim and INDUS, as well as crim and AGE. I think some or most of these are very explainable, e.g. we should expect a higher level of nox in an industrial area. If we were to take this much further we would need to do some “feature selection” to determine which variables to keep and which to set aside for various types of analyses.

Check for **linearity** by considering the residuals, specifically the residuals versus predicted values plot. This might be called any number of names. In Excel it is the X Variable Line Fit Plot. I’m pasting a copy of mine below.

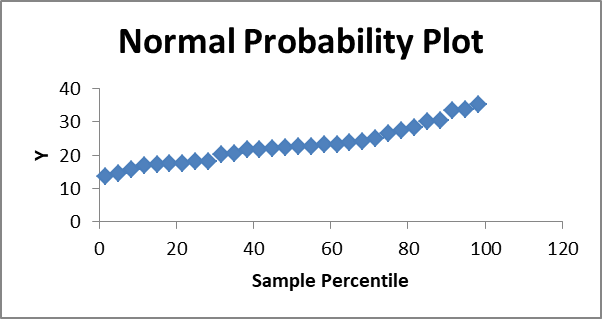


Again, this looks pretty good. For comparison, below I’m pasting one from an article at <https://www.godatadrive.com/blog/basic-guide-to-test-assumptions-of-linear-regression-in-r> related to this exercise below. I think you can immediately see the difference.

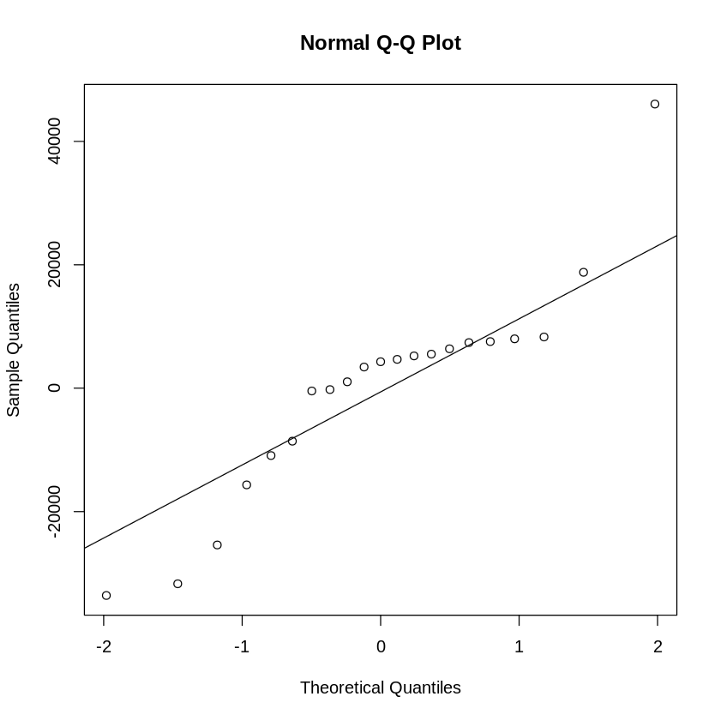


Check for **normality** by plotting. So far we’ve only considered frequency plots. You can use these to visually confirm if the plots appear normal in shape? A better way is to generate and look at a “normal probability plot” or “Q-Q plot”. This is standard output for an OLS model. It is a box you can check if you do the analysis using Excel. If normality is satisfied the plot will look like data points on a straight 45-degree line. If normality is not satisfied you may need to apply any of a variety of transformations that we haven’t considered yet. Right now you can only answer yes or no to this.

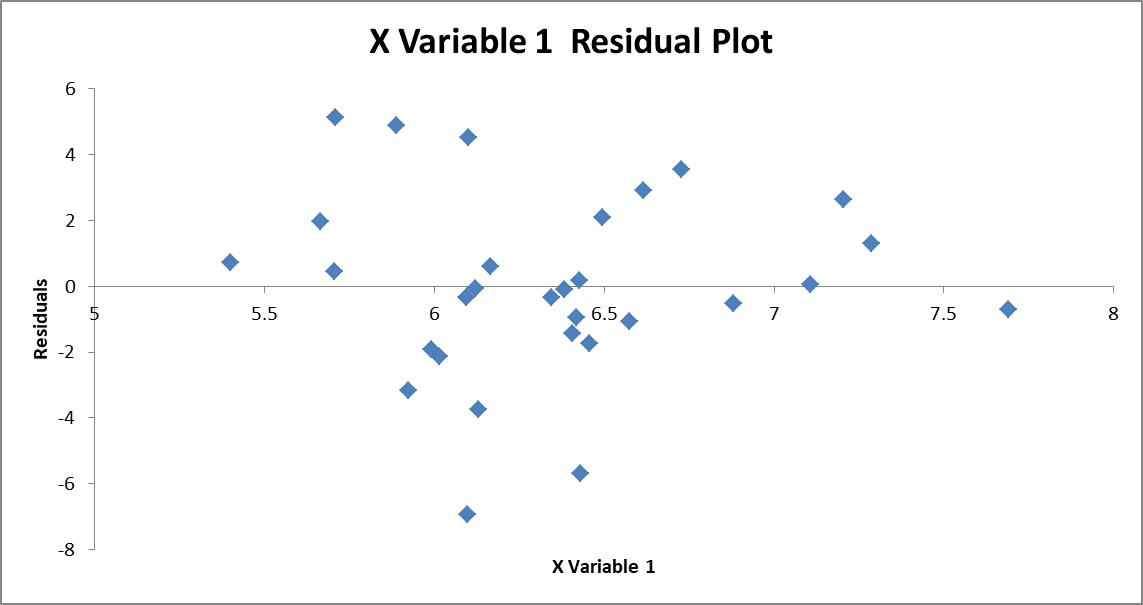
Here is my normal probability plot, often called a Q-Q plot, which doesn’t look exactly like a 45-deg line because I let Excel auto-scale the axes. But it looks pretty good anyway. Variations you might see would include wildly varying tails on the ends of the line, wild curves throughout the line, etc.



Below is a relatively bad normal probability plot from <https://towardsdatascience.com/assumptions-of-linear-regression-fdb71ebeaa8b> for comparison:



**Homoscedasticity** is checked by considering a plot of the residuals. Excel calls this the “X Variable Residual Plot. I’m pasting a copy of my plot below. This one doesn’t look as good as the others.



If I put lines along the upper and lower extremes (as in the figure below) I see what looks like a funnel.



This indicates that I have outliers in my data. The fix is to remove the outliers. However, a word of caution is in order. Before removing data you want to make sure that the outliers are not an indication of something important to your analysis. If so, you will need to consider other things to do, e.g. transformations or perhaps a different type of model such as a weighted least squares model!

The purpose of the last part of this question is not to require you to do anything but rather to introduce you to options you will have to conduct your analyses. That is, there are programs that are easier to do all this in than Excel. For example, R/RStudio which was developed specifically for statistical analyses makes this very easy. Here is what is required to do this for the same data in R/RStudio.

Commands in R/RStudio are the same blue as links (usually) are. It takes just seven commands to do everything in R/RStudio. This is why it is important to learn how to use a variety of codes. Here is my R/RStudio script including the summary of the OLS model which matches Excel’s output:

> smallHousingSample <- read.csv("I:/My Passport Documents/McDaniel/DataAnalytics/ANA500/Module3/Week5/smallHousingSample.csv")

> View(smallHousingSample)

> attach(smallHousingSample)

> model1 <- lm(CMEDV ~ RM)

> summary(model1)

Call:

lm(formula = CMEDV ~ RM)

Residuals:

Min 1Q Median 3Q Max

-6.9391 -1.3489 -0.0927 1.7978 5.1344

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -38.695 6.567 -5.893 2.44e-06 \*\*\*

RM 9.701 1.031 9.408 3.64e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.869 on 28 degrees of freedom

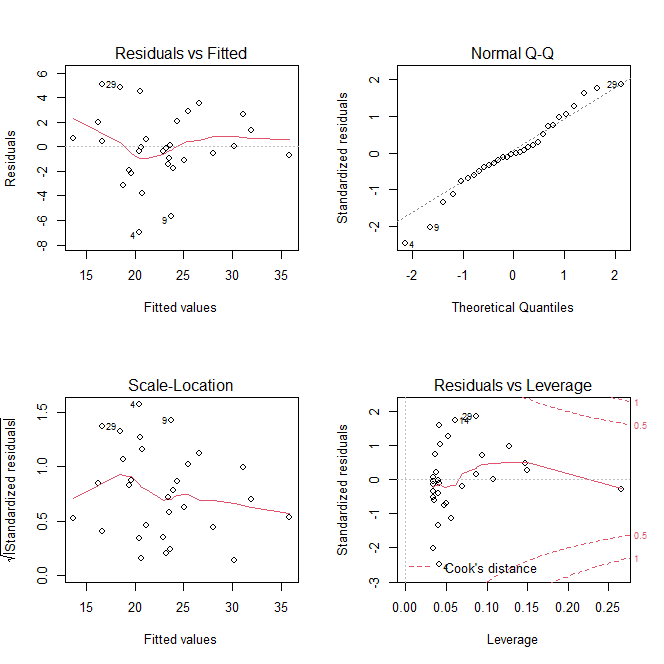
Multiple R-squared: 0.7597, Adjusted R-squared: 0.7511

F-statistic: 88.5 on 1 and 28 DF, p-value: 3.641e-10

> par(mfrow=c(2,2))

> plot(model1)

Here are the resulting plots:



Unfortunately, these plots don’t look as good as the ones from Excel even though it is all the same data! It likely has to do with the scales that each plot has.

1. Using the 30 observations sample dataset you obtained from your gretl assignment, consider the relationships between a number of the variables. For example, in the gretl assignment we considered the effect of the average number of rooms in a home to the home’s value. That is a pretty straight forward relationship. Now let’s consider some relationships that are not that straight forward. For example, is there an effect on the rate of crime due to the presence of lower socio-economic citizens?

In the case of the relationship between home value and number of rooms we have the dependent and an independent variable respectively. Now, we are considering the relationship between two of the independent variables. If we find such a relationship exists, i.e. it is statistically significant etc., then there are a number of ramifications of that relationship in conducting a regression analysis**. Let’s start by looking to see if that relationship exists to begin with. We can start with computing the correlation coefficient.** **Use the values I have computed and shared with you above to answer the following questions.**

* 1. The correlation coefficient between the rate of crime represented by the variable crim and the percentage of lower socio-economic statius population represented by the variable lstat is \_\_\_\_\_\_\_\_\_?
  2. The value of this correlation coefficient indicates that a strong relationship exists between the variables crim and lstat. True or False?

1. Again, using the output for the OLS model I generated and shared above, what is the r-squared value computed?
2. Does this r-squared value indicate that the OLS model explains most of the variability in the data?

I hope this has helped you see several things:

1. The sample size does make a difference. This is true but only up to a point. At some point you will have enough data that the model(s) you generate using it all get the same answer. After that, adding more data does not help or make a difference.
2. If you use the same data and are conducting the same analysis, e.g. generating an OLS model, you should get the same answer every time, the exception can be very small differences due to rounding.
3. There is much more to ordinary least squares than is on the surface. We’ll continue looking at regression and will introduce logistic regression as we continue our course.
4. It is important to learn how to use a variety of tools (codes) to make things easier for yourself. After all, as they say you can use a row boat to cross the Atlantic ocean but it will take a lot longer than a Cruise Ship will.