*Module 3, Week 1, gretl Problem Set 5*

This week your gretl assignment will be automated too. In this document I am providing you complete information to finish it. Once you have this gretl assignment completed then it should be easy for you to do the Paper and Pencil assignment!

You can use the following commands (between the lines below) to bring the data required to complete this assignment into gretl and create a small sample of it for the Paper & Pencil assignment for Week 5. This process eliminates observations with erroneous values. I was only able to get all variables for each observation in this way. Of course, if you know a different way to do this that’s fine too. This is just so you don’t have to think too much about importing and exporting data for your assignments right now!

**I am pasting my entire script** at the end of this document so you will have all commands in one place! That is so you do not need to try to piece a script together. You only need to read through the text of this document so you can answer the appropriate questions.

Note, in order for this to work you need to insert the complete path to the folder or directory you are opening data files from and saving your gretl data files, output, etc. to. These are the paths recovering the data file from my Midterm Exam folder and saving the newly processed data files to Module 3\Week 5\...

#Open the entire corrected Boston housing dataset

open "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Midterm\correctHousing.gdt"

#Sample and store the records before those from Boston itself

smpl 1 356

store "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\first.gdt"

#Restore the full dataset to get the remaining records

smpl --full

#Sample and store the remaining records

smpl 489 506

store "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\second.gdt"

#Open the file with the first set of records

open "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\first.gdt"

#Append the records in the second file to those in the first file

append "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\second.gdt"

#Store the subset of records

store "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\housingWOBoston.gdt"

#Plot the records with those having average rooms=0.0 omitted

gnuplot RM CMEDV --output=display

#Look at your data! You can use descriptive stats to do this

summary

#Create a small dataset for your Paper and Pencil Assignment this week. Keep in mind

#what a small sample size is!

smpl 30 --random

store "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\smallHousingSample.csv"

smpl --full

#Now that you have all the data sorted out continue with your gretl assignment

You will continue to use the gretl dataset you used for the Midterm Examination to answer the following questions.

You are interested in the relationship between the average number of rooms in owner-occupied homes (RM) and the (corrected) median home value (CMEDV) from the Boston housing dataset we used for the Midterm Examination. All answers to this gretl assignment involve your analysis of that dataset.

1. Which of the variables of interest, RM and CMEDV, are the following?
   1. The independent variable
   2. The dependent variable
2. Once you have pre-processed the data (e.g. as above between the lines) so that you have a good set of observations in your dataset, produce a scatterplot that shows how home values change with increasing number of rooms in the home.

**gnuplot** CMEDV RM --output=display

The relationship between the variables CMEDV and RM appears to be linear.

1. Calculate descriptive statistics for the variables of interest (CMEDV and RM).

**summary** RM CMEDV

What is the mean value of CMEDV?

What is the mean value of RM?

What is the interquartile range of CMEDV?

What is the standard deviation of RM?

1. Calculate the correlation coefficient between the variables of interest.

**corr** CMEDV RM

What is the value you got for the correlation coefficient?

The correlation coefficient is a measure of the strength of the relationship between two variables. True or False?

1. Residuals are the difference between the values of independent variables at different points in time. True or False?
2. The least squares method of regression to find the line best fitting the data minimizes the (select the best answer below):
   1. Sum of squared residuals
   2. Sum of residuals
   3. Sum of the dependent variable squared
   4. Sum of the difference of independent variables squared.
3. What are the assumptions required for conducting a linear regression. Select all that apply.
   1. Linearity, i.e. the relationship between dependent variable and independent variable(s) is linear.
   2. Homoscedasticity, i.e. residuals are roughly equal and scattered about zero.
   3. Independence, i.e. observations are independent.
   4. Normality, i.e. the residuals are normally distributed.
   5. The correlation coefficient between variables equals zero.
   6. The standard deviations of the dependent variable vary over time resulting in heteroscedasticity.
   7. At least one of the independent variables depends on other independent variables.
   8. The number of observations is small and the data follow a Student’s t-distribution.
   9. There are a number of explainable outliers in the data resulting in heteroscedasticity.
4. Estimate a simple linear regression model using least squares using the OLS command in gretl.

Let’s take a look at the output from the OLS command:

**ols** CMEDV 0 RMDon’t forget to include the “0” in the command for ordinary least squares computation because that is what gives you the value for the intercept.

Model 1: OLS, using observations 1-374

Dependent variable: CMEDV

coefficient std. error t-ratio p-value

---------------------------------------------------------

const −26.0455 2.23293 −11.66 5.26e-027 \*\*\*

RM 7.99551 0.349031 22.91 4.52e-073 \*\*\*

Mean dependent var 24.70294 S.D. dependent var 8.387353

Sum squared resid 10884.86 S.E. of regression 5.409286

R-squared 0.585176 Adjusted R-squared 0.584061

F(1, 372) 524.7651 P-value(F) 4.52e-73

Log-likelihood −1161.036 Akaike criterion 2326.072

Schwarz criterion 2333.921 Hannan-Quinn 2329.189

Just for comparison sake here is the output from Excel for the same data:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |  |
| Multiple R | 0.765121883 |  |  |  |  |  |  |  |  |
| R Square | 0.585411497 |  |  |  |  |  |  |  |  |
| Adjusted R Square | 0.584294007 |  |  |  |  |  |  |  |  |
| Standard Error | 5.414980483 |  |  |  |  |  |  |  |  |
| Observations | 373 |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |  |
| Regression | 1 | 15360.72425 | 15360.72425 | 523.8632123 | 6.31553E-73 |  |  |  |  |
| Residual | 371 | 10878.46706 | 29.32201364 |  |  |  |  |  |  |
| Total | 372 | 26239.19131 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |  |
| Intercept | -26.05394429 | 2.235350016 | -11.65542045 | 5.80305E-27 | -30.44948918 | -21.6583994 | -30.44948918 | -21.6583994 |  |
| 6.575 | 7.997913702 | 0.349436095 | 22.88805829 | 6.31553E-73 | 7.310789973 | 8.685037431 | 7.310789973 | 8.685037431 |  |
|  |  |  |  |  |  |  |  |  |  |

The results from Excel are virtually identical to those from gretl. Both methods’ results match!

* 1. The estimated regression equation is . True or False?
  2. Discuss the estimated slope coefficient.
     1. The value of the estimated slope coefficient is \_\_\_\_\_\_\_.
     2. Interpret the estimated slope coefficient by considering the following statement.

The estimated slope coefficient tells you how much the dependent variable, in this case home value, varies with changes in the independent variable, i.e. the average number of rooms in owner-occupied homes. True or False?

* 1. Is the estimated slope coefficient statistically significant? How do you know?
     1. yes
     2. In this case, the P-value equals 4.52e-073 \*\*\* or something very, very small and much smaller than the designated 0.05 level of significance.

1. The coefficient of determination or r-squared, is a measure of how much of the variability in the data is explains the response, i.e. the dependent variable.
2. The value of r-squared for our current model is \_\_\_\_\_\_\_\_\_?
3. At what point or value is the coefficient of determination or r-squared considered a strong indicator?
   1. 0.50
   2. 0.70
   3. 0.90
   4. It depends…

The truth is that a good value for r-squared depends on what the model you are developing is intended to do. If the model is intended to represent a lot of engineering or technical applications then usually somewhere between 0.50 and 0.70 is considered good. However, if you are developing a model for a final consumer product where safety is involved you’ll want a much higher r-squared, e.g. 0.90 or even a lot higher than that. Most basic R&D projects are good with an r-squared value of around 0.2. In this case, r-squared is only intended to give enough confidence to refine something to the next step or phase which should have a higher r-squared. In the social sciences, r-squared = from 0.10 to 0.30 is often considered good. So, it depends.

1. Calculate a 95% confidence interval for the estimated slope coefficient.

First, be sure to keep the result you want clearly in mind, i.e. you want the 95% confidence interval for the slope coefficient. We have almost all the information required to obtain the upper and lower bounds and hence the confidence interval from the OLS output. The slope coefficient is given in gretl is given by $coeff(RM). The standard error of the slope coefficient is given by $stderr(RM). The degrees of freedom is given by $df.

Using an online calculator, or other appropriate means, knowing the degrees of freedom you can get t for the confidence interval. For a 95% confidence interval the t is 2.262.

The basic way we get a lower bound, e.g. a lower bound is using the regression model (equation) to minimize the sum of squared errors by subtracting appropriate values for the lower bound. Note that it is easy to use gretl’s “critical” function which you can find more info on in the gretl function reference. The basic format of the equation for the lower bound is:

Lower bound = $coeff(RM) – t-cricical \* standard error(RM)

To get the critical value use gretl’s built-in “critical” function. For more info on that see the gretl function reference available through the help menu. The equations for the lower and upper bounds are:

scalar lb = $coeff(RM) - critical(t, $df, 0.025)\*$stderr(RM)

scalar ub = $coeff(RM) + critical(t, $df, 0.025)\*$stderr(RM)

print lb ub

* 1. The value of the lower bound for the 95% confidence interval of the slope coefficient is \_\_\_\_\_\_\_\_\_?
  2. The value of the upper bound for the 95% confidence interval of the slope coefficient is \_\_\_\_\_\_\_\_\_?

**The complete gretl Script for the Module 3 Week 5 gretl assignment.**

Note, in order for this to work you need to insert the complete path to the folder or directory you are opening data files from and saving your gretl data files, output, etc. to. These are the paths recovering the data file from my Midterm Exam folder and saving the newly processed data files to Module 3\Week 5\...

#Open the entire corrected Boston housing dataset

**open** "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Midterm\correctHousing.gdt"

#Sample and store the records before those from Boston itself

**smpl** 1 356

**store** "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\first.gdt"

#Restore the full dataset to get the remaining records

**smpl** **--full**

#Sample and store the remaining records

**smpl** 489 506

**store** "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\second.gdt"

#Open the file with the first set of records

**open** "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\first.gdt"

#Append the records in the second file to those in the first file

**append** "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\second.gdt"

#Store the subset of records

**store** "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\housingWOBoston.gdt"

#Look at your data! You can use descriptive stats to do this

**summary**

#Create a small dataset for your Paper and Pencil Assignment this week. Keep in mind

#what a small sample size is!

**smpl** 30 **--random**

**store** "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Module3\Week5\smallHousingSample.csv"

**smpl** **--full**

#Now that you have all the data sorted out continue with your gretl assignment

#

#

#Plot the records with those having average rooms=0.0 omitted

**gnuplot** CMEDV RM **--output**=display

#

#

#

#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 we can use the corresponding linear regression (line) to predict what a median home value

#would be for any number of rooms. For example, for 5 rooms we would get:

**scalar** CMEDV\_hat = **$coeff(const)** + **$coeff**(RM)\*5

#

#

#Now let's look at the 95% confidence intervals

**scalar** lb = **$coeff**(RM) - **critical**(t, **$df**, 0.025)\***$**stderr(RM)

**scalar** ub = **$coeff**(RM) + **critical**(t, **$df**, 0.025)\***$stderr**(RM)

**print** lb ub