*Module 4, Week 1, gretl problem set 7*

**Learning Outcomes:**

* Interpret model output(s)
* Work in consistent units of measure.

Because the scripts we’re writing are getting more complicated I’m pasting the entire script for this assignment at the end of this Word doc. I’ll also paste the R/RStudio script after that. If you are not familiar with R, note that the extension on R scripts is “.R” rather than “.inp” as is true for gretl.

It is just best practices to setup a directory structure for data and scripts that is separate from other types of files. For those of us who use scripts a lot you may have already setup a directory for data files, a working directory, separate directories for gretl scripts and R scripts, etc. Once you’ve been working with data and different scripts for a while you’ll figure out how you work easiest and be able to move from directory to directory saving modified data files, scripts, etc.

**gretl Problem**

Let’s start with a relatively simple situation. This involves the sales of dinners in a cafeteria. There are two independent variables; price and adv\_cost, and one dependent variable sales. The variables price stands for the price per meal and adv\_cost stands for advertising costs. We’ll let price be and adv\_cost be . Sales (revenue) will be represented by “y”.

Even this simple situation can lead to a number of questions that can be analyzed by developing a regression model. Those questions might include, “Because a number of different meals are available in the cafeteria, what is the best way to represent the overall price?” or “How does this price vary from cafeteria location to location?” Assume that there are 75 observations in the data set at 75 cafeterias across five cities. The values given are 1,000’s of dollars (USD) annually.

1. Considering the descriptive statistics for the cafeteria.gdt data, this dataset does NOT have sufficient observations (or records) to build an OLS model without worry about sample size. Therefore, we will have to modify the distributions used for tests of the model to compensate.
   1. True
   2. False
2. Thinking about consistent units and the listed units for sales, price and advertising costs (adv\_cost), an adjustment may need to be made to values in price for a direct comparison to the other variables sales (revenue) and advertising cost (adv\_cost).
   1. True
   2. False
3. The simple linear regression model for this situation is:
   1. True
   2. False

I have included tests for normality, linearity, and homoscedasticity in the script at the end of this Word doc. You can read more specifics about them in the output from the gretl script as well as in the various manuals and command help files available for gretl.

1. Using the data in the data file cafeteria.csv, consider the descriptive statistics, frequencies, and develop a simple linear model to represent these data.
   1. The assumption of normality is satisfied.
      1. True
      2. False
   2. The assumption of linearity is satisfied.
      1. True
      2. False
   3. The assumption of homogeneity or homoscedasticity is satisfied.
      1. True
      2. False

To be thorough, at this point I would like to give you the results for the same data used to build (essentially) the same model on which the same tests were conducted but in R/RStudio rather than gretl. Here is the model:

> summary(Model1\_cafe)

Call:

lm(formula = sales ~ price + adv\_cost, data = cafeteria)

Residuals:

Min 1Q Median 3Q Max

-13.4825 -3.1434 -0.3456 2.8754 11.3049

Coefficients:

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

(Intercept) 118.9136 6.3516 18.722 < 2e-16 \*\*\*

price -7.9079 1.0960 -7.215 4.42e-10 \*\*\*

adv\_cost 1.8626 0.6832 2.726 0.00804 \*\*

---

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

Residual standard error: 4.886 on 72 degrees of freedom

Multiple R-squared: 0.4483, Adjusted R-squared: 0.4329

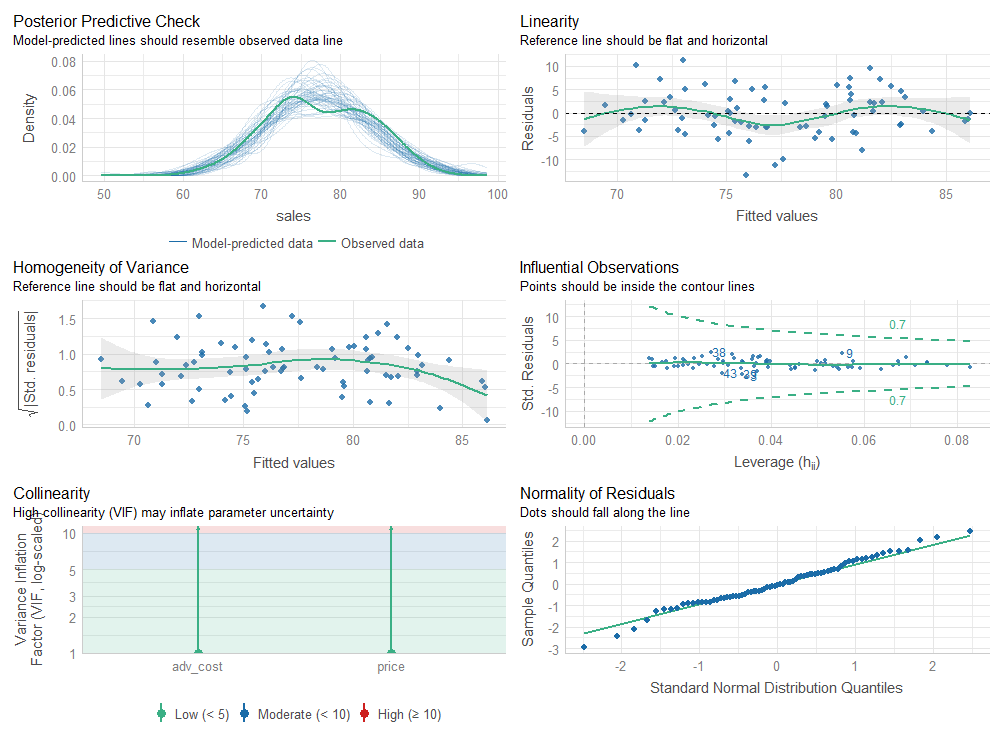
F-statistic: 29.25 on 2 and 72 DF, p-value: 5.041e-10

which is interesting because this model indicates that all variables are significant, as did the model produced by gretl, but NOT at the same level of significance (see the P-value for adv\_cost).

Below are the test results which can easily be viewed graphically in R/RStudio. The comparable check for normality is the Q-Q Plot which reveals some bit of variation but overall doesn’t look too bad. This certainly does not look as bad as the P-value of 0.60972 from gretl would suggest.

I would question linearity given the graphical results from R/RStudio which gretl output suggests is ok based on a P-value of 0. 0102601.

The gretl output for the last test, i.e. for heteroscedasticity or heteroscedasticity, had a P-value of 0.247035 indicating heteroscedasticity was present. In this case the graphical output from R/RStudio and the numerical assessment from gretl seem to agree in that the graph of “Homogeneity of Variance” is certainly not a “straight line”.



1. Analyze the coefficients related to the explanatory (independent) variables.
   1. All the variables in the model are statistically significant.
   2. None of the variables in the model are statistically significant.
   3. Only the price variable is statistically significant.
   4. Only the advertising costs variable is statistically significant.

Use the gretl OLS model output to answer the following questions.

1. The value of the intercept of the regression line is \_\_\_\_\_\_\_\_\_.
2. The coefficient of price is \_\_\_\_\_\_\_\_\_.
3. The coefficient of adv\_cost is \_\_\_\_\_\_\_\_\_.
4. How much annual revenue will be generated if the average price per meal is $5.50 and the advertising costs are capped at $1,200? In this case, be sure to round correctly to the dollar (USD). (Hint: you’ve done this before when considering other variables and ensuring you are using consistent units, i.e. remember that the problem statement says that values are in 1,000’s of dollars (USD).)
5. Determine the number of degrees of freedom for this problem. (Hint: remember the formula where N is the number of observations and K is the number of unknown coefficients.) The number of degrees of freedom is \_\_\_\_\_\_\_\_\_.
6. The 95% confidence interval for price has a
   1. lower limit of \_\_\_\_\_\_\_\_\_ and,
   2. an upper limit of \_\_\_\_\_\_\_\_\_.
7. Based on the computed 95% confidence interval this means that if the price per meal is “decreased” by 10 cents or 0.10 that means that sales (revenue) would be increased (Hint: remember that the units are 1,000’s of dollars (USD).)
   1. between a lower limit of\_\_\_\_\_\_\_\_\_\_, and
   2. an upper limit of \_\_\_\_\_\_\_\_\_\_.
8. In this question you will consider the significance of the coefficients given, for example, a decrease in price of $0.10. Assume . (Note: if you use a script to do this be very careful. The placement of the parentheses in the equations for the t-ratios will change the computed values.)

Evaluating the significance of individual coefficients to determine whether or not a variable should be included in a model is typically done by hypothesis testing using standard null and alternate hypotheses. For example to test the coefficient of price, the null and alternate hypotheses are

Even though we have more than 30 observations we still use a t-test to find a t-statistic because we do not know the population standard deviation. From the t-table,

The computed t-statistic (or t-ratio) for price is \_\_\_\_\_\_\_\_\_\_.

therefore we reject the null hypothesis and conclude that there is sufficient evidence that we should keep price in the model.

Once you have developed a script to test the significance of the coefficients, then you can modify the equations to easily conduct additional hypothesis tests for one-tail alternatives, to evaluate how much elasticity exists, etc.

1. Using the same hypothesis test or computing with gretl, the t-ratio for the advertising costs (adv\_cost) is \_\_\_\_\_\_\_\_ .
2. These values are the same as the values computed and included in the table for the original OLS model.
   * 1. True
     2. False

Because a decrease in price results in an increase in sales (revenue), that indicates that demand is elastic. If there were no increase in demand, i.e. no increase in sales with a decrease in price, then demand would be inelastic!

1. The corresponding P-values indicate that we must reject the null hypothesis, i.e. the P-value is much less than 0.05. Therefore, both price and adv\_cost are significant variables and should be included in the model.
   1. True
   2. False
2. The null and alternate hypotheses to test if any additional dollars of advertising will generate additional sales is:
   1. True
   2. False
3. Assume a level of significance of and find the relevant critical value. Treat the null hypothesis as an equality, i.e. (because that will be the limiting value) and find the critical value.
   1. The critical value for is \_\_\_\_\_\_\_\_\_\_.
4. The corresponding P-value is \_\_\_\_\_\_\_\_\_.
5. Compute any additional required values to determine whether or not there is sufficient evidence to confirm that additional dollars spent on advertising will be cost effective, i.e. result in increased sales (revenue). Answer the question, “There is sufficient evidence to confirm that additional money spent on advertising will result in increased revenue.” Yes or No.
   1. Yes
   2. No

To continue this further we would start considering whether or not there are some more complex interactions, e.g. if dropping the price of a meal if more effective than increasing advertising. Rather than considering each coefficient independently, to determine if this is true we would need to consider a linear combination of coefficients. For now, we’ll stop here. Next time we’ll continue and look at how our model might become non-linear, i.e. if variables have higher order effects or interact with each other, e.g. quadratic or cubic polynomial models, log-linear models or interaction models.

Script for Week 7 gretl Assignment:

#

#Script for gretl Week 7 Assignment

#

#My path to data file is

# "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\gretl\dataFiles\POE5\dataFiles\cafeteria.csv"

#Your path will be different!

set verbose off

#

# Example 5.1

#Explore whether or not there are sufficient observations not to worry

#about adjusting the distribution for testing for a small sample size.

#The number of observations in the cafeteria.gdt dataset is:

#

scalar obs1 = $nobs

printf "\n The number of observations in the cafeteria.gdt data set is: %d.\n",obs1

printf "\n"

printf "\n"

#

#

#Open the datafile and consider some variable units and descriptive statistics.

#

#

open "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\gretl\dataFiles\POE5\dataFiles\cafeteria.csv"

#

#

#Change the descriptive labels and graph labels

setinfo sales --description="Monthly sales revenue ($1000)" \

--graph-name="Monthly Sales ($1000)"

setinfo price --description="Price in dollars" --graph-name="Price"

setinfo adv\_cost --description="Monthly adv\_cost Expenditure ($1000)" \

--graph-name="Monthly adv\_cost ($1000)"

#

#

# print the new labels to the screen

labels

#

# summary statistics

summary sales price adv\_cost --simple

#

#Build the OLS model for the cafeteria data

#

m1<-ols sales const price adv\_cost --vcv

#

# Predict sales when price is 5.50 and adv is 1200

scalar yhat = $coeff(const) + $coeff(price)\*5.50 + $coeff(adv\_cost)\*1.2

printf "\nPredicted sales when price=$5.50 and adv\_cost=$1200 is $%.2f\n", yhat\*1000

#

#I have included tests for normality, linearity and homoscedasticity below

#You can read through the output to see the results of these tests.

#

#All these tests work in an analogous to hypothesis testing although each

#arrives at its respective P-values in a very different way.

#

#Basically, if the P-value is greater than the level of significance, e.g. 0.05, then we

#fail to reject the null hypothesis that there is no \_\_\_\_ and must assume that \_\_\_\_ is

#present in the data where \_\_\_\_ is normality, linearity or heteroskedasticity.

#

#For example, results of the Breusch-Pagan test for normality tell us whether to accept

#the null hypothesis that there is no heteroskedasticity in the data or to fail to accept

#the null hypothesis because there is heteroskedasticity present

#

#

modtest --normality

modtest --logs

modtest --breusch-pagan

#

#Consider the 95% confidence interval for price

#

printf "\n Compute a 95%% confidence interval for the variable price.\n"

printf "\n"

scalar bL = $coeff(price) - critical(t, $df, 0.025) \* $stderr(price)

scalar bU = $coeff(price) + critical(t, $df, 0.025) \* $stderr(price)

printf "\nThe lower = %.2f and upper = %.2f confidence limits\n", bL, bU

#

#

#Now test the significance of the coefficients. If a coefficient equals

#zero, then that variable should not be included in the model.

#

#

#rebuild the basic OLS Model

#

ml<-ols sales const price adv\_cost --vcv

#

#Also, we can use hypothesis testing to evaluate elasticity or lack

#thereof in demand.

#

#From the t-table our critical value is -1.666.

#

#Therefore if the test statistic is less than -1.666 or the P-value

#is less than 0.05 then we reject the null hypothesis.

#

scalar t1=($coeff(price)-0)/$stderr(price)

pvalue t $df t1

scalar t2=($coeff(adv\_cost)-0)/$stderr(adv\_cost)

printf "\n"

printf "\n"

printf "\n The t-ratio for H0: b1=0 is = %.3f.\n\

The t-ratio for H0: b2=0 is %.3f.", t1, t2

printf "\n"

printf "\n Remember that b0 is the const term.\n"

printf "\n"

printf "\n"

#

#

#Test to see if additional advertising will generate additional revenue

#Again, the critical value is -1.666 and the P-value is 0.05

#

printf "\n The P-value or \n"

scalar t3 = ($coeff(adv\_cost)-1)/$stderr(adv\_cost)

pvalue t $df t3

printf "\n The t-ratio H0: b2=1 is = %.3f.\n",t3

#

#