ANA 500 Week 8 gretl and P&P Assignments

Fall I 2022 – Dr. Hamner Sections F22A and F22C

Data have been acquired about the Baton Rouge, LA housing market are available in the batRouge.csv data file. This week’s gretl and P&P assignments comprise several sections that assess your knowledge about ANA 500 Foundations of Data Analytics. As has been the case with previous gretl and the paper and pencil homework assignments, this week’s assignments will have many different types of questions; true/false, multiple choice, multiple answer, computed numeric, etc. You should strive to select the choice that best answers the question or enter your computed numeric values rounded to two decimal places.

Parts I, II, and III will constitute your Week 8 gretl assignment. Parts IV and V will constitute your Week 8 Paper and Pencil assignment. If you want to complete these assignments using gretl’s built in menus that is fine. I am providing a complete script at the end of this document in several parts as appendices. Each appendix will be labeled with the corresponding parts of the assignments that the script pertains to. Because of the complexity of the overall script you will probably want to run this as each script in each appendix separately. You’ll want to copy your gretl output to a file, either in Notepad or Word, as well as the plots/graphs generated.

# Part I: Descriptive and Summary Statistics

Start by considering your dataset. Make a record of your answer to each of these questions, anyway you want to make this record, to use in answering later questions on this exam.

* How many observations are there?
* How many variables are there?
* What is the “type” of each variable?

Generate a table of summary statistics for your data.

* Are there any missing observations?
* Does it look like you will have to deal with any inconsistent units?

1. Generate descriptive statistics for the variables price, sqft, and age. Use these statistics to answer the following questions.
   1. The price variable is:
      1. Numeric, discrete
      2. Numeric, continuous
      3. Logical
      4. Categorical
   2. The values in the price variable are in:
      1. Dollars (USD)
      2. 10’s of dollars (USD)
      3. 100’s of dollars (USD)
      4. 1,000’s of dollars (USD)
   3. The values in the sqft variable are in:
      1. 1’s
      2. 10’s
      3. 100’s
      4. 1,000’s
   4. The values in the age variable are in:
      1. 1’s
      2. 10’s
      3. 100’s
      4. 1,000’s
   5. Based on the answers above, no adjustment or transformation should be required to interpret the results of analyses using these variables. (True/False)
2. Restrict your data to traditional-style houses. Consider descriptive and summary statistics for your restricted dataset. Use the restricted dataset to answer the following questions.
   1. How many observations are there?
   2. The correlation between traditional-style house prices and size is statistically significant. (True/False)
   3. The value of the correlation coefficient is \_\_\_\_\_\_\_\_\_.
3. Create a scatter plot of house price versus house size for traditional style homes. Does the relationship between price and size appear to be linear? (Yes/No)
4. Save your reduced dataset to a new data file, e.g. batonRouge-trad.gdt.
5. Generate a histogram of house prices. Use this to answer the following questions.
   1. Is the data skewed?
      1. No apparent skew
      2. Right skew
      3. Left skew
      4. Apparent uniform distribution
   2. Based on your answer about any apparent skew do you believe you may have to transform your data to meet the assumptions required to build a regression model? (Yes/No)
6. Generate a histogram of the natural log of house prices, i.e. . Use this to answer the following questions.
   1. Now is the data skewed?
      1. No apparent skew
      2. Right skew
      3. Left Skew
      4. Apparent uniform distribution
   2. Based on your answer about any apparent skew after taking the natural log of the price variable, do you believe you may have to further transform your data to meet the assumptions required to build a regression model? (Yes/No)
   3. Which assumption could be violated?
      1. Linearity
      2. Normality
      3. Homoscedasticity
      4. Independence
7. Create a scatter plot of the natural log of house price versus house size for traditional style homes that are owner occupied. Is the relationship between price and size linear now? (Yes/No)

# Part II: Simple Linear and Non-linear Regression

1. Generate a simple linear model for traditional style houses with price as a function of house size. That is,

(Be sure to save the value for the sum of squares error (SSE) for this linear model.)

1. Interpret the estimates to answer the following questions.
   1. Is house size statistically significant? (Yes/No)
   2. How do these house prices vary with changes in size (change per square foot)?
   3. The intercept for the simple linear model is, practically speaking, realistic. (Yes/No)
2. Create a scatter plot of house price versus house size for traditional style homes including the (linear) trend line.
3. Generate a quadratic model for this situation, that is , and use this model to answer the following questions. (Be sure to save the value of the sum of squares error (SSE) for this quadratic model.)
   1. What is the intercept value?
   2. The intercept for the quadratic model is, practically speaking, realistic. (Yes/No)
   3. What is the coefficient of ?
   4. What is the marginal effect for a home with 2000 square feet of living area?
   5. What is the expected price of the 2000 square foot home?
   6. What is the elasticity of price with respect to living area for a traditional-style home with 2000 square feet of living area?
   7. Generate a scatter plot with both the linear and quadratic trend lines on it. Which seems to fit the data better?
      1. Linear fit
      2. Quadratic fit
      3. Neither the linear or quadratic fit appear to be a “better” fit than the other
   8. Generate a plot of the residuals from both the linear and quadratic models. Does homoscedasticity appear to be a problem? (Yes/No)
   9. Would this indicate that heteroscedascity or heteroskedascity is present in the data? (Yes/No)
4. Generate a log-linear model for this situation, that is , and use this model to answer the following questions. (Be sure to save the sum of squares error (SSE) for this log-linear model.)
   1. The house size in square feet is statistically significant (Yes/No)
   2. The intercept of the log-linear model is statistically significant (Yes/No)
   3. The intercept for the log-linear model is, practically speaking, realistic. (Yes/No)
   4. Visually, the \_\_\_\_\_\_\_\_\_\_ model appears to be the best fit for the data.
      1. Linear
      2. Quadratic
      3. Log-linear
      4. All models appear to be equally good fits to the data
   5. Compare the sum of squares error (SSE) for each model and select the model listed below that actually results in the least error.
      1. Linear
      2. Quadratic
      3. Log-linear
      4. All result in the same SSE
5. Ultimately, the log-linear model results in higher house prices for very large houses. (True/False)
6. Based on the results of the various tests for normality, \_\_\_\_\_\_\_\_ satisfy/satisfies the assumption of normality. (Hint: these tests are based on **the hypothesis that the data are normal to begin with**, i.e. If the P-value is < 0.05 we must reject the null hypothesis. In other words, when evaluating your results, keep in mind what it means to have a given hypothesis and the P-values you get from your results!)
   1. the simple linear model
   2. the quadratic model
   3. the log-linear model
   4. None of the models…
   5. All of the models…
7. Visually inspecting plots of residuals indicates that \_\_\_\_\_\_\_\_ satisfy/satisfies the assumption of normality.
   1. the simple linear model
   2. the quadratic model
   3. the log-linear model
   4. None of the models…
   5. All of the models…
8. Consider the plots of residuals generated in the part of your assignments. From visually inspecting the plot do the residuals appear to be relatively evenly distributed about zero? (Yes/No)

# Part III: Point Estimation, Interval Estimation and Hypothesis Testing

1. Consider the differences in value for owner-occupied houses versus vacant/rental houses. You will need to subset the full dataset by the variable owner to do this. That is, you will have one where you restrict the data to owner=1, the other where owner=0. Generate limited log-linear models including the variables price, square feet (sqft) and age; one restricted to owner-occupied houses, the other for vacant or rental houses. Use your results to answer the following questions.
   1. The mean of the price for owner-occupied houses is \_\_\_\_\_\_\_\_.
   2. The mean of the price for a vacant or rental house is \_\_\_\_\_\_\_\_.
2. Compare the frequency plots after transforming the price variable using a natural log transformation. Do the frequency plots indicate that by taking the natural log of price we have improved the normality of the distribution? (Yes/No)
3. Using the original simple linear model developed earlier for traditional-style houses, test the null hypothesis that the expected price of a 2000 square foot house is equal to or less than $120,000. Use a level of significance equal to 0.05. Use your results to answer the following questions.
   1. The upper limit of the 95% confidence interval is \_\_\_\_\_\_\_\_.
   2. The lower limit of the 95% confidence interval is \_\_\_\_\_\_\_\_.
   3. The P-value for sqft is \_\_\_\_\_\_\_\_.
4. Using the quadratic model developed earlier for traditional-style houses that are 2000 square feet in size, test the null hypothesis that the marginal effect of an additional square foot of living area is $75 against the alternate hypothesis that the effect is less than $75. Use a level of significance of 0.01. Based on the results of your hypothesis test you fail to reject the null hypothesis and conclude that for a 2000 square foot house, the marginal effect of adding a square foot of living area is less than $75. (True/False)
5. Using the quadratic model developed earlier for traditional-style houses that are 4000 square feet in size, test the null hypothesis that the marginal effect of an additional square foot of living area is $75 against the alternate hypothesis that the effect is less than $75. Use a level of significance of 0.01. Based on the results of your hypothesis test you fail to reject the null hypothesis and conclude that for a 4000 square foot house, the marginal effect of adding a square foot of living area is less than $75. (True/False)
6. Another plot of residuals was generated in the part of your assignments. The model the residuals were obtained from added the variable age to the variables used to generate the OLS model. Does it appear that adding another variable helped satisfy the assumption of regression? (Yes/No)

# Part IV: Multivariable (or Multiple Variable) Regression

1. Add the variable age to your original simple linear model to develop a multivariable linear model. (Continue as before using the restricted dataset.) Use that model to answer the following questions.
   1. Are the intercept and all coefficients statistically significant? (Yes/No)
   2. How much does a traditional-style house increase in value for each additional square foot of living space?
2. Now return to the full, unrestricted dataset. Rerun the multivariable model including size (sqft) and age. Use your results to answer the following questions.
   1. The intercept and all coefficients are NOT statistically significant. (True/False)
   2. The difference in cost for an additional square foot of living space from the traditional-style house to all houses is \_\_\_\_\_\_\_\_.
   3. The mean of house prices increases by \_\_\_\_\_\_\_\_.
   4. An increase of one year in age decreases the house price (for all houses) by \_\_\_\_\_\_\_\_.
   5. The value of the intercept for the multivariable linear model is “practically” realistic. (True/False)
3. Add higher (second) order terms for the variables sqft and age to your multivariable model. Use your results from running that model to answer the following questions. (NOTE!!! Do not round until you are ready to submit your final answer. It does make a big difference here.)
   1. The intercept and all coefficients are not statistically significant. (True/False)
   2. Estimate the marginal effects of price as a function of size, i.e. , and age, i.e., for the smallest or oldest house in the data, the largest or newest house in the data, and a 2300 square foot or a 20 year old house. To help here are the relevant equations:

The complete model is:

The derivative with respect to size (sqft) to find the marginal effect of size is:

The derivative with respect to age to find the marginal effect of age is:

* + 1. The coefficient of the sqft term is \_\_\_\_\_\_\_\_.
    2. The coefficient of the term is \_\_\_\_\_\_\_\_.
    3. The coefficient of the age term is \_\_\_\_\_\_\_\_.
    4. The coefficient of the age term is \_\_\_\_\_\_\_\_.
    5. The estimated marginal effect of sqft on price for the smallest house, 662 square feet, is \_\_\_\_\_\_\_\_.
    6. The estimated marginal effect of sqft on price for the largest house, 7,897 square feet, is \_\_\_\_\_\_\_\_.
    7. The estimated marginal effect of sqft on price for the 2300 square foot house is \_\_\_\_\_\_\_\_.
    8. The estimated marginal effect of age on price for the oldest house, 662 square feet, is \_\_\_\_\_\_\_\_.
    9. The estimated marginal effect of age on price for the newest house, 7,897 square feet, is \_\_\_\_\_\_\_\_.
    10. The estimated marginal effect of age on price for the 20 year old house is \_\_\_\_\_\_\_\_.
    11. The lower bound on a 95% confidence interval around the marginal effect of size on price is \_\_\_\_\_\_\_\_.
    12. The upper bound on a 95% confidence interval around the marginal effect of size on price is \_\_\_\_\_\_\_\_.

# Part V: Use of Indicator Variables (Section 5.10 in textbook)

1. We know that house prices are affected by the size of the house. It also seems that house prices are greater for owner-occupied houses. Let’s quantify this. There are a couple different ways we can do this. First, consider how the size of a house that is owner-occupied affects house prices. Doing this, we can estimate the marginal effect of this.

Make an OLS model for the housing data you have. Include the variables; sqft, (square feet size), bedrooms (number of bedrooms), baths (number of bathrooms), age (years), owner (=1 if owner occupied, 0 if vacant or a rental), pool (=1 is yes, 0 if no). Add an interaction term between sqft and owner, i.e. price x owner.

Generate a “log-linear” model where you use the natural log of the price, i.e. ln(price) and the normal values of all other variables. In addition, this time to check normality generate a Q-Q Plot from your results. Remember that we determined earlier that we could make the distribution for price more normal by taking the natural log of it. Use your results to answer the following questions.

* 1. The intercept and all the coefficients are statistically significant. (True/False)
  2. The coefficient of sqft\_owner is significant at the 0.05 level of significance. (True/False)
  3. The R-squared value is \_\_\_\_\_\_\_\_.
  4. Most of the variance in the data is explained by the model. (True/False)
  5. The P-value from the F-test, i.e. P-value < 0.05, indicates the utility of the model is statistically sound. (True/False)
  6. The coefficient of the variable sqft\_owner is essentially zero but does take away a bit from house prices if a house is owner-occupied. (True/False)
  7. The marginal effect of size for owner-occupied 2000 square foot houses is \_\_\_\_\_\_\_\_.
  8. The marginal effect of size for owner-occupied 2500 square foot houses is \_\_\_\_\_\_\_\_.
  9. The value of the coefficient for sqft\_owner and its sign (plus or minus) makes sense. (True/False)
  10. The transformation by taking the natural log of the price appears to have made the price distribution more normal. (True/False)

1. Looking at the marginal effect when we simply added an interaction term didn’t really work out. And, it seems that taking the natural log of price isn’t very effective either. Much earlier we did see that using a quadratic equation resulted in the lowest values for the sum of squares error (SSE). So, let’s consider the second way to look at this. Given that owner=1 if the house is owner-occupied and =0 if the house is vacant or a rental, then we can manipulate the equations to take that into account (see Section 5.10 of your textbook). We already expressed the formula for the estimated value of price with sqft\_owner as an indicator variable. Let’s look at the equations for the case when owner=1 and owner=0 and use a quadratic form.

Here is the original, complete equation with the initial interaction term:

To remove the original interaction term and build a complete quadratic model:

We need to add additional terms for the qualitative independent variable’s interaction with the other variables:

The original interaction term is in this equation too. But there are a lot of other interactions captured by considering the variable owner as an indicator variable.

When owner = 0:

Which makes sense because all terms involving “owner” just go away. Now, when owner = 1:

Now we can get the marginal effect of size when the house is owner-occupied and when it is not. That is, taking the partial derivative of these equations with respect to sqft we get:

When “owner”=0,

When “owner”=1,

Now we build the model and get the coefficients. Use your results to answer the following questions.

1. When the house is vacant or a rental, i.e. when owner = 0, the marginal effect increases for increasing house size. (True/False)
2. When the house is owner-occupied, i.e. when owner = 1, the marginal effect increases for increasing house size. (True/False)
3. The trend of increase price as house size increases makes “practical” sense. (True/False)
4. When the house is owner occupied the marginal effect of increasing size is doubled. (True/False)
5. Consider the plot of the residuals. Does it appear that they are evenly distributed about zero? (Yes/No)

Appendix A: Part I and Part II of the Assignments

#

#Script for ANA 500 Week 8 gretl and P&P assignments

#

#

#Part I Descriptive and Summary Statistics

#

#Don't forget to change the path to wherever you have stored the data file!!!!!

open \

"I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Exams\FinalFallI-2022\br2.gdt"

#

#Generate information about the data set as well as descriptive and summary statistics

#

scalar obsTotal = $nobs

printf "\n The number of observations is %d.\n",obsTotal

#

labels

#

info

#

summary

#

#Setup the data to generate the required scatter plot for traditional style homes

#

smpl traditional==1 --restrict

gnuplot price sqft --fit=linear --output=display

store "I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Exams\FinalFallI-2022\br2-trad.gdt" --gzipped=1

#

#

#Print the number of observations of tradiitonal-style homes

#

scalar obsTrad = $nobs

printf "\n The number of observations of traditional-style homes is %d.\n",obsTrad

#

#Consider the relationship between price and home size

#

corr price sqft

#

#Generate frequency plots to consider transforming the price variable

#using a natural log transformation

freq price --plot=display

series l\_price = ln(price)

freq l\_price --plot=display

#

#Generate a plot of ln(price) versus sqft

gnuplot l\_price sqft --fit=semilog --output=display

#

#Part II: Simple Linear and Non-Linear Regression

#

#

#\*\*\*\*\*\*

#Build a simple linear model

#\*\*\*\*\*\*

#

ols price const sqft --vcv

#Save the sum of squares error

sse\_l = $ess

#

#Save the residuals to plot to verify assumptions

#

series uhat1\_l = $uhat

gunplot uhat1\_l sqft --plot=display

modtest --normality

modtest --logs

modtest --breusch-pagan

#

#Compute a 95% confidence interval for this model and the associated P-value

#

printf "The number of degrees of freedom is %.2f.\n",$df

scalar tcrit = critical(t, $df, 0.025)

printf "The critical value is %.2f.\n",tcrit

#

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

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

printf "For the data restricted to traditional-style houses, \n\

the lower bound is %.2f and the upper bound is %.2f.\n",lb,ub

#

#Method to compute the P-value

#

#scalar t1 = critical(t, $df, .025)

#printf "\nThe value of t-critical is %.2f.\n",t1

#scalar t2 = ($coeff(sqft)-0)/$stderr(sqft)

#scalar pval = pvalue(t, $df, t2)

#printf "\nThe P-value is %.20f.\n", pval

#

#\*\*\*\*\*\*

#Build a second-order model

#\*\*\*\*\*\*

#

series sqft\_sq = sqft^2

ols price const sqft\_sq --vcv

#Save the sum of squares error

sse\_2 = $ess

#Save the residuals

series uhat1\_q = $uhat

#generate the corresponding scatter plot

gnuplot price sqft\_sq sqft --output=display

modtest --normality

modtest --logs

modtest --breusch-pagan

#

#\*\*\*\*\*\*

#Build a log-linear model

#\*\*\*\*\*\*

#

logs price

ols l\_price const sqft

series l\_yhat = $yhat

series ln\_pri = exp(l\_yhat)

#Compute the sum of squares error for the log-linear model

#Note that the output for sse in the log-linear case is NOT accurate

series y\_diff = (price-ln\_pri)^2

scalar sse\_ln = sum(y\_diff)

#Save the residuals

series uhat1\_ln = $uhat

#Create the corresponding plot

gnuplot price ln\_pri sqft --output=display

modtest --normality

modtest --logs

modtest --breusch-pagan

#

#Consolidate into one plot to examine the differences between models

#

gnuplot price sqft\_sq ln\_pri sqft --output=display

#Print out the quantitative differences in SSE between models.

printf "\n The Sum of Squares Error (SSE) for the linear model is %.3f\n",sse\_l

printf "\n The Sum of Squares Error (SSE) for the quadratic model is %.3f\n",sse\_2

printf "\n The Sum of Squares Error (SSE) for the log-linear model is %.3f\n",sse\_ln

#

#Verify model assumptions

#

#Use the Jarque-Bera statistic to verify normality

#Normal random variables have no skew or excess kurtosis

#Note that the JB test statistic value gets larger with increasing skew and kurtosis

#

normtest uhat1\_l --all

normtest uhat1\_q --all

normtest uhat1\_ln --all

#

#Create plots of residuals

#

gnuplot uhat1\_l sqft --output=display

gnuplot uhat1\_q sqft --output=display

gnuplot uhat1\_ln sqft --output=display

Appendix B: Part III for gretl and P&P Assignments

#

#Part III: Point Estimation, Interval Estimation and Hypothesis Testing

#

#Don't forget to change the path to wherever you have stored the data file!!!!!

open \

"I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Exams\FinalFallI-2022\br2.gdt"

#

#Consider the difference in house value when the house is owner-occupied versus

#a vacant/rental house.

#Subset the data for owner=1

smpl owner==1 --restrict

#Build a log-linear model

#

series l\_price = ln(price)

ols l\_price const sqft age

#Save the residuals

series uhat1 = $uhat

#Save the coefficients

matrix b = $coeff

b

#Save the standard errors

matrix serr = $stderr

serr

series l\_yhat = $yhat

series ln\_own = exp(l\_yhat)

#Compute the sum of squares error for the log-linear model

#Note that the output for sse in the log-linear case is NOT accurate

series y\_diff = (price-ln\_own)^2

scalar sse\_ln = sum(y\_diff)

#Check some assumptions

#Plot the frequencies to see if we have a more normal distribution

#for price using l\_yhat. Compare agains a freq plot of price

freq price --plot=display

freq l\_yhat --plot=display

#Plot the residuals

gnuplot uhat1 sqft --output=display

#Create the corresponding plots of results

#In the first plot l\_yhat still has not been transformed for final output

gnuplot price l\_yhat sqft --output=display

gnuplot price ln\_own sqft --output=display

#Predict the price for a 2000 square foot house that is 20 years old

#and is owner-occupied

scalar pri\_owner\_2000 = exp(($coeff(const) + $coeff(sqft)\*2000 + $coeff(age)\*20))

pri\_owner\_2000

#Retrieve the full dataset again

smpl full

#Now subset the data for owner=0

smpl owner==0 --restrict

#Build a log-linear model

#

series l\_price = ln(price)

ols l\_price const sqft age

#Save the residuals

series uhat1\_ln = $uhat

#Save the standard errors

matrix serr = $stderr

series l\_yhat = $yhat

series ln\_rent = exp(l\_yhat)

#Compute the sum of squares error for the log-linear model

#Note that the output for sse in the log-linear case is NOT accurate

series y\_diff = (price-ln\_rent)^2

scalar sse\_ln = sum(y\_diff)

#Generate a Q-Q Plot to check normality

qqplot uhat1\_ln --output=display

#predict the price for a 2000 square foot house that is 20 years old

# and is vacant or a rental

scalar pri\_rental = exp($coeff(const) + $coeff(sqft)\*2000 + $coeff(age)\*20)

pri\_rental

#Create the corresponding plot

gnuplot price ln\_rent sqft --output=display

#Consolidate results in one plot

gnuplot price ln\_own ln\_rent sqft --output=display

Appendix C: Part IV: Multivariable (or Multiple Variable) Regression

#

#Part IV: Multivariable (or Multiple Variable) Regression

#

#Don't forget to change the path to wherever you have stored the data file!!!!!

open \

"I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Exams\FinalFallI-2022\br2.gdt"

#

#Build multivariable models

#

#Recover the full dataset

smpl full

#Start with a linear model

ols price const sqft age --vcv

#Build the multivariable linear model

ols price const sqft age --vcv

#Save the sum of squares error (SSE)

sse\_lmulti = $ess

#Save the residuals

series uhat2\_l = $uhat

#

#Add second order terms to the multivariable linear model to build a complete

#quadratic model. Remember that before you only added the one second-order term

#to create a somewhat limited quadratic model

#

series sqft\_sq = sqft^2

series age\_sq = age^2

ols price const sqft sqft\_sq age age\_sq --vcv

#Save the sum of squares error

sse\_2 = $ess

#Save the standard error of regression

se\_reg = $sigma

#Save the residuals

series uhat1\_q = $uhat

#Save the matrix of coefficients and se

matrix b = $coeff

matrix cova = $vcv

matrix serr = $stderr

#

#Compute the 95% confidence interval for the marginal effect around size

#for a 2300 square foot house

siz = 2300

serr\_sqft = ((serr[2,]^2)+(((2\*siz)^2)\*(serr[3,]^2))+(2\*(2\*siz)\*cova[2,3]))^0.5

print serr\_sqft

der\_sqft = b[2,]+(2\*b[3,]\*siz)

print der\_sqft

scalar tcrit = critical(t, $df, .025)

lb\_multi2 = der\_sqft - (tcrit\*serr\_sqft)

ub\_multi2 = der\_sqft + (tcrit\*serr\_sqft)

printf "\nThe 95%% Confidence Interval for the marginal effect of size on price \

is from $%.2f to $%.2f.\n",lb\_multi2,ub\_multi2

#

#reopen the data file to eliminate rewriting over variables

open \

"I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Exams\FinalFallI-2022\br2.gdt"

#

#Generate the interaction term

genr sqft\_owner = sqft \* owner

#

#Build the Log-linear model

logs price

ols l\_price const sqft bedrooms baths age owner sqft\_owner

series l\_yhat = $yhat

series ln\_price = exp(l\_yhat)

#Save the residuals

series uhat1\_ln = $uhat

#Setup the equations to get the results to consider the effect of size on owner-occupied

#houses

scalar me2000 = $coeff(sqft)+2000\*$coeff(sqft\_owner)

scalar me2500 = $coeff(sqft)+2500\*$coeff(sqft\_owner)

printf "\nThe marginal effect of size for owner-occupied 2000 square\

foot houses is %.2f.\n",me2000

printf "\nThe marginal effect of size for owner-occupied 2500 square\

foot houses is %.2f.\n",me2500

#Create the corresponding plot

gnuplot price ln\_price sqft --output=display

qqplot ln\_price --output=display

freq price --plot=display

freq ln\_price --plot=display

Appendix D: Part V: Use of Indicator Variables (Section 5.10 in the textbook)

#

#Part V: Use of Indicator Variables (Section 5.10 in textbook)

#

#

#reopen the data file to eliminate rewriting over variables, etc.

open \

"I:\My Passport Documents\McDaniel\DataAnalytics\ANA500\Exams\FinalFallI-2022\br2.gdt"

#

#First setup all the required terms and then

#Build the quadratic model

series sqft\_sq = sqft^2

series bedrooms\_sq = bedrooms^2

series baths\_sq = baths^2

series age\_sq = age^2

series owner\_sq = owner^2

series owner\_sqft = owner\*sqft

series owner\_bedrooms = owner\*bedrooms

series owner\_baths = owner\*baths

series owner\_age = owner\*age

series owner\_sqft\_sq = owner\*sqft\_sq

series owner\_bedrooms\_sq = owner\*bedrooms\_sq

series owner\_baths\_sq = owner\*bedrooms\_sq

series owner\_age\_sq = owner\*age\_sq

series owner\_owner\_sq = owner\*owner\_sq

#

ols price const sqft bedrooms baths age owner sqft\_sq bedrooms\_sq baths\_sq age\_sq owner\_sq \

owner\_sqft owner\_bedrooms owner\_baths owner\_age owner\_sqft\_sq owner\_bedrooms\_sq owner\_baths\_sq \

owner\_age\_sq owner\_owner\_sq

series l\_yhat = $yhat

series ln\_price = exp(l\_yhat)

#Save and plot the residuals

series uhat1\_ln = $uhat

gnuplot uhat1\_ln sqft --output=display

#Setup the equations to get the results to consider the effect of size on owner-occupied

#houses

#When owner = 0

scalar me2\_2000 = $coeff(sqft)+2\*$coeff(sqft\_sq)\*2000

scalar me2\_2500 = $coeff(sqft)+2\*$coeff(sqft\_sq)\*2500

#When owner = 1

scalar me3\_2000 = ($coeff(sqft)+$coeff(sqft))+2\*($coeff(sqft\_sq)+$coeff(sqft\_sq))\*2000

scalar me3\_2500 = ($coeff(sqft)+$coeff(sqft))+2\*($coeff(sqft\_sq)+$coeff(sqft\_sq))\*2500

#

#When owner = 0, the house is vacant or a rental

printf "\nThe marginal effect of size for owner-occupied 2000 square\

foot houses is %.2f.\n",me2\_2000

printf "\nThe marginal effect of size for owner-occupied 2500 square\

foot houses is %.2f.\n",me2\_2500

#

#When owner = 1, the house is owner occupied

printf "\nThe marginal effect of size for owner-occupied 2000 square\

foot houses is %.2f.\n",me3\_2000

printf "\nThe marginal effect of size for owner-occupied 2500 square\

foot houses is %.2f.\n",me3\_2500