ANA 510 Problem Set #1

Fall II 2022

Pretend you are part of a large real estate firm in Ultimate City named Ultimate Realty. Your job is to study the real estate market and provide information to the real estate agents and brokers in the firm in order to support sales goals. The Office Manager is your boss. The tools you have available are not the most sophisticated. However you believe that with the knowledge you have gained in the Data Analytics program, you can do a good job for the firm. Ultimate Real Estate currently has two brand new customers, Mr. Road Runner and Mr. Wylie Coyote. Mr. Road Runner has a house for sale. Mr. Wylie Coyote is currently looking for a property to purchase (possibly with nefarious intentions).

Data acquired about the Ultimate City housing market are available in the ultimateHousing.csv data file. You will use this data to complete this assignment to help your review of the material covered in ANA 500 Foundations of Data Analytics. Be sure to record your answers to questions in a secondary location (notepad) or Word doc. In case you have an unanticipated power failure or internet interruption. In these cases I can reset your exam to allow you to start over. However, that reset process will wipe out any work you have already completed.

As has been the case with gretl and paper and pencil homework assignments, you can use the tools you want to use to complete it. As usual, this examination will have many different types of questions; true/false, multiple choice, multiple answer, computed numeric, etc. You should strive to select the choice or choices that best answer(s) the question or enter your computed numeric values rounded to two decimal places, unless specifically instructed otherwise.

NOT ALL QUESTIONS BELOW WILL BE INCLUDED IN THIS ASSIGNMENT IN BLACKBOARD. OR, SOME OF THE QUESTIONS MIAY BE WORDED SLIGHTLY DIFFERENTLY. SO, BE SURE TO READ THE QUESTIONS IN BLACKBOARD CAREFULLY TO SEE WHICH YOU NEED TO ANSWER AND WHAT THE CORRECT ANSWER SHOULD BE!

# Part I: Descriptive and Summary Statistics

I suggest that you begin in the usual way by getting to know your data.

1. How many observations about the real estate market in Ultimate City have been acquired?
2. What explanatory data (independent variables) are available?
3. How many categorical variables are available?
4. Does it **initially** look like you will need to make any adjustments to account for differences in units of measure, e.g. from data in 1,000’s to data in 1’s? (Yes/No)
5. Are any of the variables in the data highly correlated? (Yes/No)
6. If you answered yes to Question 5, which variable would you consider dropping first?
   1. I answered “no” to Question 5
   2. Bedrooms
   3. Baths
   4. Lgelot
   5. Pool
   6. Age
7. The data contain two variables that might be used as indicator variables; lot size (lgelot) and pool.
   1. What proportion of the data is lgelot? (leave your answer as a decimal number not percent)
   2. What proportion of the data is pool? (Leave your answer as a decimal number not percent)
8. Is the distribution of your dependent variable normal? (Yes/No)
9. If you answered no to Question 8 what transformation would you probably use to help force compliance with the assumption of normality?
   1. I answered “yes” to Question 8
   2. Quadratic
   3. Square root
   4. Inverse (1/variable)
   5. Ln(variable), i.e. take the natural log of the variable

# Part II: Regression

1. Consider the different kinds of models we have studied in ANA 500, e.g. simple linear, multivariable linear, non-linear either simple or multivariable, etc. Which of the following is the type of test you would use to determine which type of model is best?
   1. Goodness-of-Fit Test
   2. Test of Independence
   3. Comparing Sum-of-Squares-Error (SSE) to find the least error
   4. Test for Homogeneity
   5. The test I would use is not in this list of tests
2. Test different models to determine which model you should use to compute the information Ultimate Realty needs. That is, consider the differences in the sum-of-squares-error values, frequency plots to check normality, scatter plots to check linearity, etc. From the choices below select which model you have found is best overall.
   1. Simple linear
   2. Simple quadratic (non-linear, polynomial)
   3. Simple log-linear (non-linear, logarithmic)
   4. Simple log-quadratic
   5. Multivariable linear
   6. Multivariable quadratic
   7. Multivariable log-linear
   8. Multivariable log-quadratic
   9. All the models tested will provide the same accuracy and quality of results
   10. None of the models tested will provide results
3. As often happens, regardless of what you think is best your boss decides that you will use a multivariable log-quadratic model as follows. Use the variables price, size (sqft), size squared (sqft\_sq), age, age squared (age\_sq), bedrooms, and baths. And, he/she says for you to take the natural log of the price variable, i.e. ln(price).

Answer the following questions based on results from that model.

* 1. Except for the ends of the plot, the Q-Q plot generally indicates that the assumption/condition of normality has been met. (Yes/No)
  2. The plot of residuals versus fits indicates that the assumption/condition of linearity has been met. (Yes/No)

1. Build and use the model your boss wants to answer the following questions.
   1. Consider the intercept and all coefficients of this model to select the best choice(s) from the list below.
      1. The intercept and all coefficients are statistically significant.
      2. Neither the intercept nor any of the coefficients are statistically significant.
      3. The variable, baths, is not statistically significant.
      4. There is no way to tell anything about statistical significance from the results.
   2. This model explains most of the variance in the data. (Yes/No)
   3. Based on the results there is “utility” in this model. (Yes/No)
2. The R-squared value indicates that very little of the variance in the data is explained by the model. (True/False)

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

1. The boss has convinced you that this is a pretty good model for Ultimate Realty. Because the boss is particularly interested in how size and age affects house prices evaluate those parameters and use the results to answer the following questions.
   1. Hypothesis tests can be conducted using a t-statistic and/or its related P-value. If the t-statistic falls in the rejection or the P-value is less than the level of significance of the test then the null hypothesis must be rejected. The alternative, for example a test to establish that a relationship between an explanatory (independent) variable and a response (dependent) variable exists is accepted. (True/False)
   2. Test the coefficient of size to determine if size is related to price. Based on a hypothesis test using the null hypothesis, i.e. the coefficient equals zero. There is not sufficient evidence (statistical significance) that a relationship exists between the explanatory variable and the response (dependent) variable. Therefore, size can be removed from the regression model without any impact on the results. (True/False)
   3. Does the age of a house really affect its price? Again based on a hypothesis test, the coefficient of age equals zero indicating that there is no relationship between the explanatory variable, age, and the response (dependent) variable price. (True/False)
   4. One of Ultimate’s sales agents has just spoken with Mr. Road Runner about listing his house. The house is 10 years old, has 2000 square feet of living area (sqft), and 3 bedrooms. How much should Ultimate’s agent tell Mr. Road Runner he should list his house for, i.e. the price?
   5. What is the upper bound for a 95% confidence interval in dollars (USD) based on the information in part d above?
   6. What is the lower bound for a 95% confidence interval in dollars (USD) based on the information in part d above?
   7. If Mr. Road Runner increase his living room with 200 more square feet of living space, how much would that increase the price of the house?

# Part IV: Use of Indicator Variables

1. The boss would like to look at some different results. He/she says it is because there are two explanatory variables that have not yet been explored, large lot size (lgelot) and pool.

To answer this question, consider the coefficients in this new model. Now, all coefficients are statistically significant. (True/False)

1. How much more, in a percentage, is the price of a house on a large lot (greater than 0.5 acres) than a house on a small lot?
2. What is the percent increase in price for an additional 100 square feet of living space if the house is NOT on a large lot (holding all other variables constant)?
3. What is the percent increase in price for an additional 100 square feet of living space if the house is on a large lot?