STAT 650 Assignment-06

Due: November 08, 2022 5:59PM

Instructions :

- This assignment is based on materials coverved in Lectures 15, 16, and 17.
- We highly recommend that you write your solutions in **Jupyter Notebook** and convert them to a **PDF** file. However, you may write the solutions by hand, scan and upload it as **.pdf** file.
- The PDF file should be under 15MB in size. It must be uploaded as a single file and not separate files for separate pages. Do not take a photo of each page and then paste them into a document - this will make your file too big and the results will generally not be very readable anyway.
- Please make sure that the solutions are neat, legible and in order (even if you choose to solve them in different order).
- Include **STAT650--UIN** at the top of the first page.
- Name the file as **UIN_assign6.pdf** (For eg, if someone's UIN is 123456789, then the file should be named 123456789_assign6.pdf). Otherwise, your submission will not be graded.
- You should upload your file through Canvas. You can make multiple submissions within the deadline, but only the latest submission will be considered for grading.
- You may take 6 hours extra after the due time, but 10% of your marks will be deducted.
- It is strictly prohibited to share or distribute the content in this document.

The aim of this assignment is to get familiar with Hypotheses testing.

Question 1

The **dataFile6_Q1.csv** contains data concerning pavement durability. There are measurements of the change in rut depth (y) of 31 experimental asphalt pavements that were prepared under different conditions specified by the values of five explanatory variables The variables are as follows

1. x_1: viscosity of asphalt,
2. x_2: percentage of asphalt in surface course,
3. x_3: percentage of asphalt in base course,
4. x_4: percentage of fines in surface course,
5. x_5: percentage of voids in surface course,
6. x_6: an indicator variable that results for 16 pavements tested in one set of runs from those 15 tested in the second run.

An objective of this experiment is to determine the important variables that affect the change in rut depth.

1. Create a dataframe named **df** by loading **dataFile.csv** into Python.
2. Examine the relationships (linear or non-linear) between five explanatory variables and the dependent variable, the change in rut depth (using appropriate graphical interpretations).

3. Using scatter plots, examine the relationship between the change in rut depth (y) and asphalt viscosity (x_1) in their original scale (i.e. x_1, y) and log scale (i.e., $\log x_1, \log y$). Describe briefly the importance of transforming x_1 and y into log scale.

4. Create a new dataframe (**df_trans**) by replacing the variable x_1 and y with their log transformed values. Also update the names of x_1 and y as **LogX1** and **LogY**, respectively.
5. Fit the regression model given below using the **statmodels** package and the **df_trans** dataframe.
$$LogY = \beta_0 + \beta_1 LogX1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$
6. Identify the explanatory variables that do have a significant impact on explaining the change in rut depth (use $\alpha = 0.05$).
7. Update your model by including the significant explanatory variables found in step 6. Examine whether the residuals follow the normal assumption.
8. Do you think it is necessary to include an interaction term between the variables **LogX1** and x_6 (i.e., $LogX1 * x_6$) into the model that you created in step 4 in order to improve model performances? Please provide sufficient evidence to support your answer.

(**Important:** Please do not print entire Python outputs in your answers. It is critical that you include only the information necessary to answer the question. Marks will be deducted if you include unnecessary information.)

Question 2

The dataset used for this question is named **dataFile6_Q2.csv** and is related to the **Anscombe's quartet** which illustrates the various forms of misalignment between data and models. The **dataFile2.csv** file consists of four datasets ($\{x_1, y_1\}, \{x_2, y_2\}, \{x_3, y_3\}$ and $\{x_4, y_4\}$).

1. Load **dataFile2.cvs** into Python and create a dataframe named **df**.
2. Analyze which descriptive statistics are approximately similar across the four datasets.
3. Using the following Python code, fit regression models for each dataset.
   a. `import statsmodels.formula.api as smf.`
   b. `smf.ols(formula= y ~ 1 + x, data=df).fit().`
4. Create a dataframe named **Models** which includes intercept (β_0), slope (β_1), coefficient of determination (R^2), and predicted y values ($\hat{y}$) of the four models created in step 3.
5. Show the actual data and fitted line on the same graph and display four graphs in a 2×2 graph.
6. Consider the second dataset (i.e., {x2,y2}). Use the Python function `np.poly1d(np.polyfit(x, y, order))` to fit a polynormial regression line that accurately describes the data. On the same graph, plot the actual data and the predicted line.
7. Identify outliers in the dataset using the model fitted to the third dataset, $\{x_3, y_3\}$. Fit a linear model to data that is free of outliers. On the same graph, plot the actual data and the predicted line.

(**Important:** Please do not print entire Python outputs in your answers. It is critical that you include only the information necessary to answer the question. Marks will be deducted if you include unnecessary information.)

Question 3

In this question, we use the dataset **dataFile6_Q3.csv** that is comprised of records from 109 different models of cars under five parameters:

1. **cyl:** Number of cylinders,
2. **origin:** Car origin, 1 = US; 2 = Europe; 3 = Asia, an integer vector,
3. **turn:** Turn diamater, a numeric vector,
4. **hp** Horsepower, a numeric vector,
5. **mpg** Miles per gallon in city driving, a numeric vector.

Load dataFile3.csv into Python as a dataframe named df and answer the following questions.

1. Explore the relationships between these five variables using the pairplot function available in the seaborn package.
2. Investigate the relationships (linear or non-linear) between five variables (you may use a heatmap to illustrate the relationships).
3. Suppose you want to explore variability in *mgp* in response to change in **cyl, origin, turn** and **hp** by using the linear model
$$mgp = \beta_0 + \beta_1 cyl + \beta_2 origin + \beta_3 turn + \beta_4 hp$$
. Use **statmodels** package and the **df** dataframe to find the model parameters.
4. What are the most significant variables that contribute to the variability in miles per gallon in city driving? Print model summary excluding variables that are not significant from your model.
5. Utilize appropriate graphical illustrations to examine the normality of residuals.
6. After loading the **mistat** package, use the following codes to perform stepwise (or forward) regression.
   a. `outcome = 'mpg', all-vars = ['cyl', 'origin', 'turn', 'hp']`
   b. `included, model = mistat.stepwise-regression(outcome, all-vars, df)`
7. Do you see a difference between the model performance obtained in step 4 and 5?

(**Important:** Please do not print entire Python outputs in your answers. It is critical that you include only the information necessary to answer the question. Marks will be deducted if you include unnecessary information.)

Question 4

In this question, you use the dataset **dataFile6_Q4.csv** that is comprised of records from 768 different people to determine whether a given patient shows sign of diabetes. The dataset consists of information of eight attributes and a label to indiate a patient or a healthy individual:

1. **pgt:** Number of times pregnant
2. **Gls:** Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. **bp:** Diastolic blood pressure (mm Hg)
4. **skt** Triceps skin fold thickness (mm)
5. **si:** 2-Hour serum insulin (mu U/ml)
6. **bmi:** Body mass index (weight in kg/(height in m)^2)
7. **dpf:** Diabetes pedigree function
8. **age:** Age (years)
9. **y:** Class variable (0 or 1)

Load dataFile4.csv into Python as a dataframe and answer the following questions providing only necessary Python outputs. Please do not to include (or print) codes (or outputs) that are not required to answer the questions.

(**Note:** Marks will be deducted for unclear and unnecessary Python codes and Python outputs that do not have sufficient explanations.)

1. Briefly describe relationships between the nine attributes listed above using pairplots and correlation plots (i.e., Pearson, Spearman and Phik correlation matrices as heatmaps).
2. Include attributes 1-8 and 9 in two dataframes named X and y.
3. Create training (i.e., X_train, y_train) and testing (X_test, y_test) datasets from the dataframes X and y in step 3, allocating 60% of samples to training set and 40% to testing set.
4. Apply logistic regression to the training dataset and evaluate classifier performance by computing the accuracy score and confusion matrix with the testing dataset (set the number of iterations to 1000).

5. By using the **MinimaxScaler** python function, scale the X matrix created in step 2. Next, repeat steps 3 and 4 to employ the logistic regression with the training sample size set to 80% and the number of iterations set to 1000. Do data scaling and increasing training data size contribute to improved classification performance?
6. To perform cross-validated logistic regression, use the function **LogisticRegressionCV(cv, random_state, max_iter)** and the data used in step 5 (i.e., scaled X). The cv value is set to 10 (i.e., 10-fold cross-validation), the random_state is set to 0, and the maximum iteration is 1000. Using the accuracy score and confusion matrix, evaluate the performance of the classifier.
7. To answer this question, use the classification model created in step 6. Calculate the ROC curve using the steps provided below.

   a. Define 20 threshold values as `threshold = np.linspace( 0, 1, 20 )`,

   b. Use `y-pred = ( model.predict-proba( X-test )[ :, 1 ] >= threshold[ i ] ).astype( int )`, to predict y values (y_pred) for each threshold value (i.e., threshold[i], where $i = 0, 1, \cdots, 19$) and **model** is the classifier name used in step 6),

   c. Use the confusion matrix to calculate the sensitivity and specificity for each threshold value,

   d. Plot Sensitivity vs 1- spefifcity. Plot a line connecting the points (0,0) and (1,1) on the same graph.