STAT 5000 LAB #10
FALL 2024 DUE TUE, NOV 19TH NAME: _____

Directions: Complete the exercises below. When you are finished, turn in any required files online in Canvas, then check-in with the Lab TA for dismissal.

Multiple Linear Regression in SAS

Data on new vehicles for the 2014 model year are available from the Environmental Protection Agency. A random sample of 200 vehicles was selected. Using these data, we wish to predict the CO2 emissions of the vehicles in city driving (`cityCO2`). The explanatory variables are listed below.

Engine: displacement of the engine in liters (Min = 1, Max = 6.8)

Cylinder: number of cylinders (Min = 3, Max = 12)

CityMPG: Fuel economy in city driving (MPG) (Min = 11, Max = 40)

Gears: number of gears (Min = 1, Max = 9)

Intake: Number of intake valves per cylinder (Coded as 1 if 2 and 0 otherwise)

Exhaust: Number of exhaust valves per cylinder (Coded as 1 if 2 and 0 otherwise)

The data for this example are saved in the `epaCO2.txt` file located in our course's shared folder on SAS Studio. The SAS code to analyze this data using multiple linear regression (MLR) is found in the `epaCO2_Lab10.sas` file and described below.

- First, read in the data set using a tab as the delimiter:

```
data epa;
     infile '~/my_shared_file_links/u63538023/STAT5000_Fall2024_ISU/epaCO2.txt'
           dlm="    " firstobs=2;
     input engine cylinder cityMPG gears intake exhaust cityCO2;
run;
```

- Next, you can explore the explanatory variables using the `corr` procedure to find all pairwise correlations and the `sgscatter` procedure to create a matrix of pairwise scatterplots with single variable histograms along the diagonal:

```
proc corr data=epa;
   var engine cylinder cityMPG gears cityCO2;
run;

proc sgscatter data=epa;
   matrix engine cylinder cityMPG gears intake exhaust cityCO2/
       diagonal=(histogram)
       markerattrs=(size=10 symbol=CircleFilled color=black);
run;
```

  These plots will help you determine whether the MLR model is appropriate
  (look for linear patterns and moderate to high correlation with the response variable).

- Finally, you can conduct MLR for a variety of models using the `reg` procedure:

```
proc reg data=epa;
   model cityCO2 = engine cylinder cityMPG;
   model cityCO2 = engine cylinder cityMPG gears;
   model cityCO2 = engine cylinder cityMPG gears intake;
run;
```

Assignment

1. Use SAS to run the multiple linear regression model for `cityCO2` with `Engine`, `Cylinder`, and `CityMPG` as explanatory variables. Use the output to complete the following exercises.
   (a) Give the equation for predicting the `cityCO2` values from the three explanatory variables.
   (b) Conduct an F-test for the overall model in helping to explain the `cityCO2` values. Report the null and alternative hypotheses, test statistic and p-value, and interpret the result in the context of the study.
   (c) Give the value of R^2 for this model and interpret its value (in context).
   (d) Conduct a t-test for the significance of `Engine` in the model that includes `Cylinder` and `CityMPG`. Report the null and alternative hypotheses, test statistic and p-value, and interpret the result in the context of the study.
2. Use SAS to run the multiple linear regression model for `cityCO2` with `Engine`, `Cylinder`, `CityMPG`, and `Gears` as explanatory variables. Use the output to complete the following exercises.
   (a) How much does adding `Gears` to the multiple linear regression model with `Engine`, `Cylinder`, and `CityMPG` reduce the sums of squared errors?
   (b) How much does adding `Gears` to the multiple linear regression model with `Engine`, `Cylinder`, and `CityMPG` increase the value of of R^2?
   (c) Conduct an F-test for the effect of adding `Gears` to the multiple linear regression model with `Engine`, `Cylinder`, and `CityMPG`. Report the null and alternative hypotheses, test statistic and p-value, and interpret the result in the context of the study.
3. Use SAS to run the multiple linear regression model for `cityCO2` with `Engine`, `Cylinder`, `CityMPG`, `Gears`, and `Intake` as explanatory variables. Use the output to complete the following exercises.
   (a) Give the equation for predicting the `cityCO2` values from the four explanatory variables for vehicles with two intake valves per cylinder and for vehicles that do not have two intake valves per cylinder. What is the difference in these two equations?
   (b) Conduct a t-test for the significance of `Intake` in the model that includes `Engine`, `Cylinder`, `CityMPG`, and `Gears`. Report the null and alternative hypotheses, test statistic and p-value, and interpret the result in the context of the study.
4. How do the 3 MLR models compare?

Total: 50 points **# correct:** _____ **%:** _____