WDS HW 3

Group Member 1
Group Member 2
Group Member 3
Group Member 4
Group Member 5

Due: 10:00PM, July 21, 2021

Contents

1 Case study: Automobiles efficiency **2**

2 EDA **2**

2.1 What effect does `time` have on `MPG`? 2

2.2 Bring origin into the model 2

2.3 Prediction 3

1 Case study: Automobiles efficiency

Are cars being built more efficient? Are Asian cars more efficient than cars built in America or Europe? To answer the questions we will use the `Auto` dataset from ISLR. The original dataset contains 408 observations about cars. It is similar to the CARS dataset that we use in our lectures. But it also collects information by years. To get the data, first install the package ISLR. The `Auto` dataset should be loaded automatically. The original data source is here: https://archive.ics.uci.edu/ml/datasets/auto+mpg

Get familiar with this dataset first. A good data set should be well documented. Use the command `?ISLR::Auto` to view a description of the dataset. Please add the variable list with names, brief descriptions and units of the variables below.

2 EDA

Explore the data first.

i. What is the range of `year`? Why is this important to know?
ii. Should `origin` be a continuous variable? Why or why not. In any case make `origin` a categorical variable.
iii. Do you see any peculiarity in the data?

2.1 What effect does `time` have on `MPG`?

i. Show a scatter plot of `mpg` vs. `year` with the LS line imposed. Does the plot show a positive trend?

ii. Now run a simple regression of `mpg` vs. `year` and report R's `summary` output. Is `year` a significant variable at the .05 level? State what effect `year` has on `mpg`, if any, according to this model.

iii. Add `horsepower` on top of the variable `year` to your linear model. Is `year` still a significant variable at the .05 level? Give a precise interpretation of the `year`'s effect found here.

iv. The two 95% CI's for the coefficient of year differ among ii. and iii. How would you explain the difference to a non-statistician?

v. Create a model with interaction by fitting `lm(mpg ~ year * horsepower)`. Is the interaction effect significant at .05 level? Explain the year effect (if any).

2.2 Bring origin into the model

i. Do mpg's differ on average among different origin? Fit a linear model with mpg vs. origin. Report the output.

a) Are mpg's on average different among three regions? Perform a test at .01 level. When you reject the null hypothesis, what have you proved?

b) Describe on average which origin has the highest mpg and what it is. Which origin has the smallest mpg on average and what is it?

c) Are Asian cars more efficient than American cars? Produce a 95% CI's for the difference.

ii. Try to build a final model which includes `year` and `origin`.

a) Describe the final model. Include diagnostic plots with particular focus on the model residuals and diagnoses.

b) Describe the year effect and origin effect in this final model. Are cars being built more efficiently over time? Are Asian cars more efficient than cars built in America or Europe?

2.3 Prediction

Use the final model to predict the `mpg` of the following car: A red car built in the US in 1983 that is 180 inches long, has eight cylinders, displaces 350 cu. inches, weighs 4000 pounds, and has a horsepower of 260. Also give a 95% CI for your prediction.