Lab 5: Regression
CSC 2621 Introduction to Data Science

Learning Outcomes

1. Select, apply, and interpret appropriate visual and statistical methods to analyze distributions of individual variables and relationships between pairs of variables.

2. Communicate findings through generated data visualizations and reports.

3. Determine and apply appropriate experimental setup, evaluation metrics, and models for supervised learning problems.

4. Perform and interpret feature selection to identify relationships between features and predicted variables.

Overview

For lab 6, you are going to analyze a data set called mtcars. The data set comes from a 1974 issue of Motor Trend US magazine and contains fuel consumption and 10 aspects of automobile design and performance for 32 automobiles from 1973-1974. The data set has 32 observations (rows) and 11 variables.

The columns are as follows:

- mpg -- miles per gallon
- cyl -- numbers of cylinders
- disp -- displacement (cu. in.)
- hp -- gross horsepower
- drat -- rear axle ratio
- wt -- weight (in 1000s of lbs)
- qsec -- 1/4 mile time
- vs -- Engine Shape (0 = V-shaped, 1 = straight)
- am -- transmission (0 = automatic, 1 = manual)
- gear -- number of forward gears
- carb -- number of carburetors

You will perform an exploratory data analysis (EDA) and then build a linear regression model to predict mpg.

Instructions

I want you to create your own notebook analyzing the data. In particular, you will try to identify variables that will be good predictors of the fuel economy (mpg) of the car. The basic steps will involve:

1. Load, transform, and clean the data.

- Describe the original data and transformed data using `head()`, `info()`, and `describe()`.

2. Characterize each variable.

- Plot the distributions of each variable and describe the range of values.

3. Explore the relationships between each variable and mpg.

- Choose the appropriate plots based on the variable types (e.g., categorical, numerical, ordered.)
- Which variables show statistically significant association with the response – use an alpha of 0.001)

4. Create four linear regression models:

a. A baseline or null model using only the intercept

b. A model using the variables you identified as predictive in your EDA.

c. A model built using a "greedy" approach.

- Build a model for each variable individually.
- Sort the variables using a metric of explained variance.
- Starting with the baseline model, add each variable one at a time to the model. If variable improves the model over the last model, keep that variable in the model. If the variable does not improve the error, skip that variable. At the end, you should have a single model with multiple variables.

d. A model built using a "greedy" approach using standardized values.

5. Model Evaluation and Reflection

a. Compare the Root Mean-Squared Error (RMSE) for the four models. Which model has the lowest RMSE? Which model has the highest RMSE? How do you interpret this value?

b. For the best performant model, make a scatter plot between the model's predicted mpg and the real mpg. Interpret this plot.

c. For the best performant model, plot the residuals. Does this model meet the assumptions associated with a BLUE model?

d. Using the summary table of the best performant model, which variables were the most predictive? Can you give explanations for why you think these variables were the most predictive based on your exploratory analysis?

e. Which approach (exploratory analysis, greedy, or all variables) produced the best model?

I will be will looking for following:

- Write an introduction (including your own statement of the problem), your guesses, and a written summary of your results at the top of the notebook in Markdown. Make sure to put your name at the top of the notebook.
- That you successfully imported the data and verified that the correct shape of the data
- That you properly cleaned and converted columns to appropriate types
- Used the appropriate plots to investigate the distributions of each variable.
- Used the appropriate plots to investigate relationships between the other variables and mpg.
- Evaluated your guesses in light of the evidence from the plots
- Overall, I want to see a finished, relatively polished product. Use Markdown cells in appropriate places to explain what you are doing, interpret your results, and describe your conclusions. One way to verify that your notebook is in a good state is to restart the kernel and re-run everything. This ensures that all of the necessary code is there.

Rubric

Followed submission instructions	5%
Title, Your Name, Introduction, and Conclusion	10%
Data Cleaning and Transformation	10%
Exploratory Data Analysis (EDA)	20%
Linear Regression Models	20%
Model Evaluation and Reflection	30%