Lab 8: Regression

For lab 8, 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 12 variables. We will only use the following columns:

- 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. In particular, you will try to identify variables that will be good predictors of the fuel economy (mpg) of the car.

1 Load the data

1. Describe the original data and transformed data using head(), info(), and describe().
2. Standardize your features. You may do this 'by hand' or using StandardScaler from sklearn.preprocessing.
3. Characterize the dependent variable – plot the distribution over mpg and describe the range of values.
4. Answer the following question in your writeup:
   (a) Do you need to standardize your dependent/target/response variable? Would you expect the results to change if you did?

2 Conduct Exploratory Data Analysis

1. Explore the relationships between each variable and mpg – choose the appropriate plots based on the variable types (e.g., categorical, numerical, ordered).
2. Use hypothesis testing to assess which variables show statistically significant association with the response – use $\alpha = 0.005$).
3. Answer the following question in your writeup:
   (a) Explain why we used a threshold of 0.005.

3 Create some linear regression models

1. Use LinearRegression from sklearn.linear_model to create your regression models. To do this you will need to instantiate the model, fit the model with the .fit() method, and then generate predicted values with the .predict() method. Make sure you include an intercept in all models. Do not make a test-training split of any kind. Make all model comparisons using adjusted R^2.
2. Model 1 should be a baseline or null model using only an intercept (no other features).
3. Model 2 should be a model using the variables you identified as predictive in your exploratory data analysis from section 2.
4. Model 3 should be a model built using greedy feature selection (we will discuss this further later in the term – for now, just follow the instructions). To build this model:
   (a) Build a model for each variable individually.
   (b) Sort the variables using adjusted R^2.
   (c) Starting with the baseline model, add each variable one at a time to the model. If variable improves the adjusted R^2 over the last model, keep that variable in the model. If the variable does not improve the adjusted R^2, do not keep that variable. At the end, you should have a single model with multiple variables.
5. Compute and compare the Root Mean-Squared Error (RMSE) for these three models.
6. For the best performant model, make a scatter plot between the model's predicted mpg and the real mpg.
7. For the best performant model, plot the residuals.
8. Answer the following question in your writeup:
   (a) Which model has the lowest RMSE? Which model has the highest RMSE? How do you interpret this value?
   (b) Include and Interpret the two plots you made. Does this model meet the assumptions associated with a BLUE model?
   (c) Which approach (exploratory analysis or greedy) produced the best model?

4 Assess feature predictiveness with bootstrapping

1. When you run a linear regression the impact of each feature on the dependent variable is summarized by the coefficient associated with that variable, learned by the model. In this section we will use bootstrapping to assess the empirical probability that each feature increases mpg consumption by looking at the variation of the coefficient learned by regression for each feature.

2. To do this:

   (a) Draw 32 samples from the rows of the dataset with replacement – this will be your bootstrap data set.

   (b) Instantiate a new linear regression model object. Make sure to include a constant.

   (c) Use the .fit() method to fit the linear regression to the bootstrap dataset with mpg as the target/response/dependent variable and all remaining variables as predictors.

   (d) Use the .feature_names_in_ and .coef_ methods to pull out the coefficient for each variable. Record each coefficient.

   (e) Repeat steps 1–4 1000 times. At the end you will have a list of 1000 coefficient estimates for each variable. Each one of these numbers is a measurement of how much impact this variable has on mpg.

3. Use this data to make a table that shows the empirical probability that each variable increases mpg consumption.

4. Answer the following question in your writeup:

   (a) For each variable what percentage of the time was the learned coefficient greater than 0 across the bootstrap iterations? Include a table of this information.

   (b) Which variables would you say were the most predictive? Can you give explanations for why you think these variables were the most predictive based on your exploratory analysis?

5 Submission Instructions

- Write up the answers to the questions in a short word document; aim for around 2 pages of text and include all graphics generated. Add footnotes identifying which sentence addresses which questions. Write in complete sentences organized into paragraphs – your goal is to explain what you've done and what you've learned to your audience (me!). Accordingly, you may seek to emulate some of the sections of the whale paper describing their data. Include the appropriate plots you've generated as mentioned above. Convert this to pdf and submit it. Submit your .ipynb file as well.

- The grading rubric for this assignment will be available in Canvas.

- NO OTHER SUBMISSION TYPES WILL BE ACCEPTED.
- **Late policy**: 5% of total points deducted per day for three days – after that no submissions allowed.