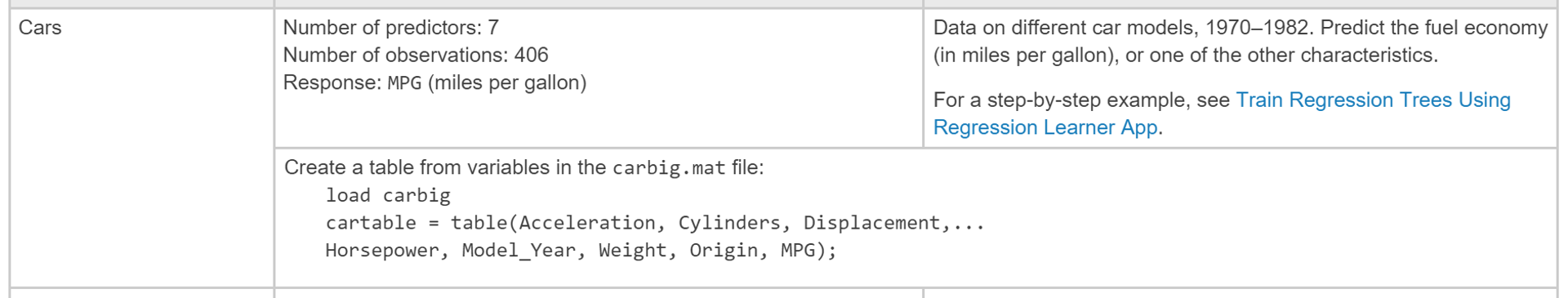
Regression Homework

Datasets are either in data folder (Exercise Datasets) or in Matlab’s demo dataset directory.

1. Use the Advertising data in Exercise Dataset
   1. Create histograms and also scatter plots of Sales vs. each of the 3 media data. A Box plot is optional (Matlab has box plot; more work to generate one in Excel)
   2. Build a linear regression model of Sales using all 3 media data as features using Excel
      1. Is this a good model?
      2. What does the coefficient and t-test’s p-value tell you?
      3. Rebuild the model without newspaper as a feature.
      4. Adding TV\*Radio as a new feature, rebuild the model and compare with the previous models.
      5. Use Matlab Regression Learner to build the model with the original data using the PCA feature with 95% variance threshold. How many principal components does it choose and what are the percent variances they represent?



1. The dataset Fish-3Only-Random in Exercise Dataset contains the weight, height, and width data of 3 different kinds of fish: Bream, Roach, and Perch. Use Excel to do the following:
   1. Since fish species is a categorical variable please do one-hot encoding for Bream and Roach (name your dummy variables Bream-D and Roach-D). Perch would then be represented by (0,0)
   2. Split the data into Train (80%) and test (20%) for each kind of fish. I have randomized the data so you can take the last 20% of each fish.
   3. Build a linear regression model for all fish’s Weight with height, width and the two one-hot variables as features. Write the formula of your model.
   4. Report the R-sqaure and RMSE of the training set. Check other key regression performance criteria.
   5. Calculate the R-sqaure and RMSE of the test set. Are they close to the training set?
   6. In the original dataset the data were sorted by fish weights (low to high). It would NOT be appropriate to take the last 20% of each fish as test data without randomizing first. Why?
2. Use the CARBIG data in Matlab demo dataset and Matlab’s Regression Learner.



* 1. Explore the data. Note that there are several variables that are not numerical in nature but for simplicity you do not need to use them for the model.
  2. Create a table of the variables of selected features and mpg as the output variables (no need to report your table in your report).
  3. Create the correlation matrix.
     1. How correlated are the feature variables?
     2. Based on the level of correlations which features do you think you need to predict mpg? Note that Regression Learner now has a Stepwise Linear option.
     3. Do you need to transform or add higher order terms to your features?
  4. Build the best MPG model you can using linear regression. Optionally you can try other algorithms.
  5. Use the PCA feature (95% threshold) to re-do the regression and compare with your model. How many Principal Components are needed and what are the % explained variance.