NYC DATA SCIENCE ACADEMY

Machine Learning Lab: Building Bridges

Data Science Bootcamp

Outline

- Part 1: Data Background
- Part 2: Initial Exploratory Data Analyses
- Part 3: Data Modeling
- Part 4: Investigative Task

PART 1

Data Background

Data Background

- Whenever a bridge is built, the project tends to go through many stages of production. One of the initial phases is mainly concerned with design. It would be helpful to predict the design time of a project in order to account for both internal and external scheduling purposes.

- The `[05] Bridges.txt` dataset includes six different variables:
  - Time: The design time in person-days.
  - DArea: The deck area of the bridge (in square feet).
  - CCost: The construction cost.
  - Dwgs: The number of structural drawings produced.
  - Length: The length of the bridge.
  - Spans: The number of spans on the bridge.

PART 2

Initial Exploratory Data Analyses

Initial Exploratory Data Analyses

1. Read the data into R and provide some basic numerical summary information on each variable.

2. Temporarily create two new categorical variables. (**Hint:** Use the `which()` function with the `%in%` operator, or use the `cut()` function) :
   a. One variable that bins the number of structural drawings into groups of 3 - 5 drawings, 6 - 8 drawings, and > 8 drawings.
   b. One variable that bins the number of bridge spans into groups of 1 span and > 1 span.

3. Based on the groups you created in part 2, are the number of structural drawings independent of the number of bridge spans? Perform and interpret the appropriate hypothesis test.

Initial Exploratory Data Analyses

4. Assess the extent of missingness in the original dataset:
   a. Visually.
   b. Numerically.

5. Impute any missing values in the original dataset using K Nearest Neighbors (use the $\sqrt{n}$ heuristic for your choice of K). Use this dataset for the the Data Modeling section of this lab.

PART 3

Data Modeling

Data Modeling

1. Create a scatterplot of bridge design time versus the number of drawings.

2. Regress time onto the number of drawings and assess the fit of the model.

3. Consider a Box-Cox transformation of your data and refit the model. Assess the fit of the transformed model.

4. Write out the equation of the transformed model and interpret the meaning of the slope coefficient in context of the problem.

5. Fit a saturated model regressing the log of design time onto all other variables in the model (on their original scale). Assess the fit of the model and discuss any outliers/influential observations.

Data Modeling

6. What does the BoxCox procedure indicate for the model you created in part 5?

7. What do the variance inflation factors indicate for the model you created in part 5? Verify your conclusions graphically using a scatterplot matrix and added variable plots.

8. Run stepwise regression to refine your model. Assess the model fit and compare this model to the saturated model you created in part 5. Which would you pick?

9. Fit a new model that predicts the log of design time by the number of drawings and the number of spans; this time, treat the number of `Spans` as a factor (use the `as.factor()` function). Interpret this model and compare it to the model you chose in part 8; which would you choose?

PART 4

Investigative Task

Investigative Task

1. Using the `cars` dataset, regress the distance the car travels onto the speed of the car (you do not need to assess the model fit). The equation should be of the following form:

$$Distance = \beta_0 + \beta_{speed} Speed$$

Note that we can rearrange this equation as follows:

$$Speed = -\frac{\beta_0}{\beta_{Speed}} + \frac{1}{\beta_{Speed}} Distance$$

2. The second equation looks very similar to the structure of a regression of speed onto distance (as if we flipped the axes). Based on the fitted regression of part 1, calculate and store the new intercept and slope of the second equation. **Do not actually perform a second regression.**

Investigative Task

3. Create a scatterplot of speed versus distance. Add to the plot:
   a. The line as defined by the intercept and slope you calculated in part 2.
   b. The actual regression line as defined by regressing speed onto distance.

4. Would you expect these lines to be the same or different? Why? What exactly is going on here?