NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp Fall 2015

Multiple Linear Regression

Question #1: Multiple Linear Regression on New York City Restaurants

Load the `04NYCRestaurants.txt` dataset into your workspace. This dataset contains survey results from customers of 168 different Italian restaurants in the New York City area. The data are in the form of the average of customer views on various attributes (food, decor, and service) scored on a scale from 1 to 30, along with the average price of dinner. There is also a categorical variable for the location of the restaurant.

1. Create a scatterplot matrix of all continuous variables colored by Location. From this plot alone, do you see any problems that might arise for multiple linear regression?
2. Fit a multiple linear regression predicting the price of a meal based on the customer views and location of the restaurant. For this model:
    a. Write out the regression equation.
    b. Interpret the meaning each of the 5 coefficients in context of the problem.
    c. Are the coefficients significant? How can you tell?
    d. Is the overall regression significant? How can you tell?
    e. Find and interpret the RSE.
    f. Find and interpret the adjusted coefficient of determination.
3. Investigate the assumptions of the model using the `plot()` function. Are there any violations?
4. Investigate the influence plot for the model. Are there any restaurants about which we should be concerned?
5. Investigate the coefficient variance inflation factors; use these values to discuss multicollinearity.
6. Create added variable plots for this model. What conclusions might you draw from these plots?

7. Fit a new simple linear regression that predicts the price of dinner from the service rating alone. Discuss this regression in light of your answer to part 6.

Question #2: Model Selection on New York City Restaurants

Continue using the `04NYCRestaurants.txt` dataset.

1. Regress the price of dinner onto the average customer food rating, decor rating, and the restaurant location. In context of this new model, comment on:
    a. The model `summary()` output.
    b. The assumptions of multiple linear regression.
    c. The influence plot of the model.
    d. The variance inflation factors of the coefficients.
    e. The added variable plots for the model.
2. Run a partial F-test to compare this model with the overall model you created in question 1. Interpret your results.
3. Fit a new reduced model that predicts the price of dinner by only the average customer food rating and average customer decor rating. Briefly comment on the model assumptions.
4. Compare each of the following models based on AIC:
    a. The overall model fitted in question 1.
    b. The overall model without the service variable fitted in question 2 part 1.
    c. The reduced model fitted in question 2 part 3.
5. Compare each of the models based on BIC.
6. Do you expect to see the results from part 4 and part 5? Which model would you ultimately choose to use?