Homework 1

March 3, 2018

Data Analysis 1

We are interested in analyzing the data contained in the attached file (car-data.txt)

The variables in the data sets are as follows: the car's manufacturer, the suggested retail price (namely, what the manufacturer thinks the vehicle is worth, including adequate profit for the automaker and the dealer, in U.S. dollars), the dealer's cost (what the dealership pays the manufacturer, in U.S. dollars); the width (in inches), the length (in inches) and weight of the car (in pounds); the horse power; the engine size (in liters); the number of cylinders; the fuel consumption in miles per gallons in the city and on the highway; the wheelbase (the distance between the centers of the front and rear wheels); and whether the car is a hybrid (1) or not.

Specifically we wish to study the effects of different aspects of a car on its retail price.

1) Answer the question above using the following car's features: engine size, number of cylinders, horse power, the highway mpg, the weight, the wheelbase and whether it is a hybrid. Specifically, build a linear regression model with these seven variables as regressors; determine whether or not it is a valid model and check whether the model assumptions are satisfied (look at the fitted line; consider residual plots; study the leverage points and outliers; draw the marginal plots). Determine whether the variables (regressors and response) should be transformed and if so, study the transformed model. For the transformed model (or the original model, if no transformation was needed), interpret the model (for example, stating which variables are important and why, interpreting the regression coefficients in the model). Draw added variable plots.
   Using an F-test, compare the model with a smaller model without the wheelbase and the fuel consumption variables. Can we drop these two variables?

2) If one is interested in estimating the effect of a car's manufacturer on the retail price, how can the model be expanded?

Data Analysis 2

Consider the data set, nyc.cvs, which contains data on Italian restaurant in NY collected in 2004. The variables are: Price (the price in US dollars of a dinner, including one drink and the tip); Food (the customer rating of the food, out of 30); Decor (the customer rating of the décor, out of 30); Service (the customer rating of the service, out of 30) and a variable (East) indicating whether the restaurant is east (1) or west of Fifth Avenue.

We want to study the effect on the price of all other variables.

1) Build a regression model using the ratings on food, decor, service and the location as regressors. Check the model validity (following the same steps as above) and comment on the results

2) Consider adding to the model interaction terms of all regressors with the variable East. Compare this model with the previous one and state whether there is evidence to support the need for different models for the East and West

Transforming the regressors

Consider a multivariable linear model. Do the regression coefficient estimates change if the regressors are centered, and if so how? By centering we mean we shift the regressor to have mean zero

$$\tilde{\mathbf{x}}_j^T = (x_{1j}, \ldots, x_{nj}) \mapsto (x_{1j} - \bar{x}_j, \ldots, x_{nj} - \bar{x}_j)$$

with $\bar{x}_j = \sum_{i=1}^n x_{ij}$. What if we standardize each variable (that is, if we center it and then divide it by its standard deviation: $s_j = \sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2/(n-1)}$

How do the regression estimate changes if the response is centered?

After you obtain the results mathematically, you may want to check it using the function scale() (notice that by default it scales and centers the data, that is, the default values of the two options are scale=TRUE, center=TRUE).