

Topics to be covered next week

STA 137

Winter Quarter, 2021

Monday, January 11 Review of regression (Handouts 2 and 3).

Wednesday, January 13 Estimation of trend (Chap 1.5 in the text, and Handout 4).

Friday, January 15 Estimation of trend and seasonal effects (Chap 1.5 in the text, and Handout 5).

Homework 1: Due by 5pm, Sunday, January 17.

You may form a group of up to 3 students (including yourself) registered in this course and submit one completed homework for the group. The front page should display only the names of the students in the group.

You will find a data set (cars93) with information on all the models on cars sold in the US in 1993. The response variable (ie, dependent variable) is car price. Independent variables are city mileage, highway mileage, engine size, horsepower (HP), tank size and weight. Before you start, please transform price, city mileage, highway mileage and engine size by natural logarithm, and horsepower by a square root. Thus

$Y = \log(\text{price})$, $X = \log(\text{city mpg})$, $X_2 = \log(\text{hwy mpg})$, $X_3 = \log(\text{eng size})$, $X_4 = \sqrt{\text{HP}}$, $X_5 = \text{tank size}$, $X_6 = \text{weight}$.

A multiple regression model is to be applied to analyze this data set

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_6 X_{i6} + \varepsilon_i, i = 1, \dots, n = 93,$$

where $\{\varepsilon_i\}$ are independent $N(0, \sigma^2)$ variables.

1. (a) Obtain a histogram for each of the variables. Are there noteworthy features in the plots? Comment.
(b) Obtain a matrix plot of the data (ie, plot all the variables against each other (R command: pairs)). Also obtain the correlation matrix. What do the plots suggest about the nature of relationship between Y and each of the predictor variables? Discuss. Does it seem that there is a problem of multicollinearity? Explain.
(c) Fit a multiple regression model to the data. Obtain the parameter estimates, their standard errors, analysis of variance table, R^2 and R^2_{adj} .
(d) Does it seem that all the independent variables need to be retained in the regression model? If you consider deleting only one independent variable, which is the best candidate for deletion? Explain your answers.

2. The questions here are on the fitted model in (1c).

- (a) Obtain a plot of the observed against the fitted Y values. Also plot the residuals against the fitted values. Does it seem that the fitted model is reasonable? Do you suspect any nonlinearity? Is the assumption of equal variance of the errors (ie, ε_i 's) reasonable here? Explain your answers.

(c) Obtain a histogram of the residuals. Also obtain a normal probability plot of the residuals, and the correlation between the residuals and the normal scores. Is the assumption of normality of the errors reasonable? Explain.

3. (a) This question is on model selection by backward elimination using AIC or BIC criterion. Fit the final selected model, obtain the parameter estimates, their standard errors, R^2 and R^2_{adj} .

(b) Use the BIC or C_p criterion to select the best among all possible regression models. Fit the final selected model(s), obtain the parameter estimates, their standard errors, R^2 and R^2_{adj} . Compare your result with the model obtained in part (a).