R Assignment 3

Chen Ming

May 27th

2017022002

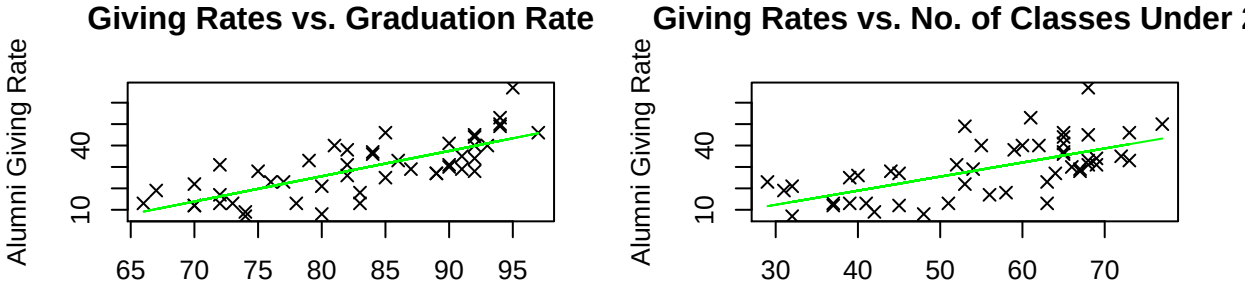
Part 1

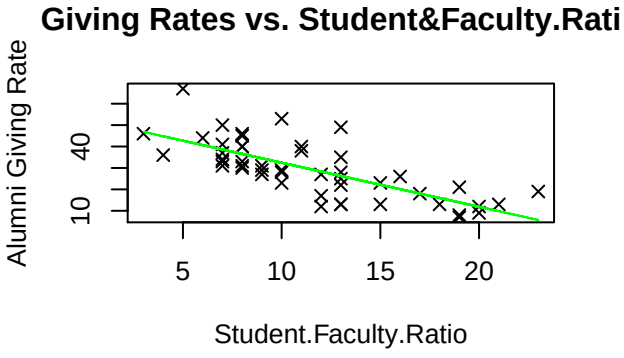
1.1. Find descriptive statistics of the data and summarize them into a table

Experiment details are omitted here.

1.2. Use graphical analysis to investigate the relationship between Alumni Giving Rate and each of the other variables

Scatter plot is demonstrated as follows. In the first and the third graphs, we can find that points cluster closely around fitted lines. While in the second graph, points seem to drift away from fitted line. From the graphical analysis, we can reasonably assume that alumni giving rate is more closely related to both graduation rate and faculty rate than number of classes under 20.





1.3.Develop a multiple linear regression model that could be used to predict the Alumni Giving Rate using the data provided

From scatter plot developed previously, we can estimate that giving rates are more closely related to **graduation rate** and **student&faculty ratio**. Therefore, we can develop the regression model by taking

the two factors into consideration.After fitting, the equation is given by:

$$y = -19.1063 + 0.7557x_1 - 1.2460x_2 \tag{1}$$

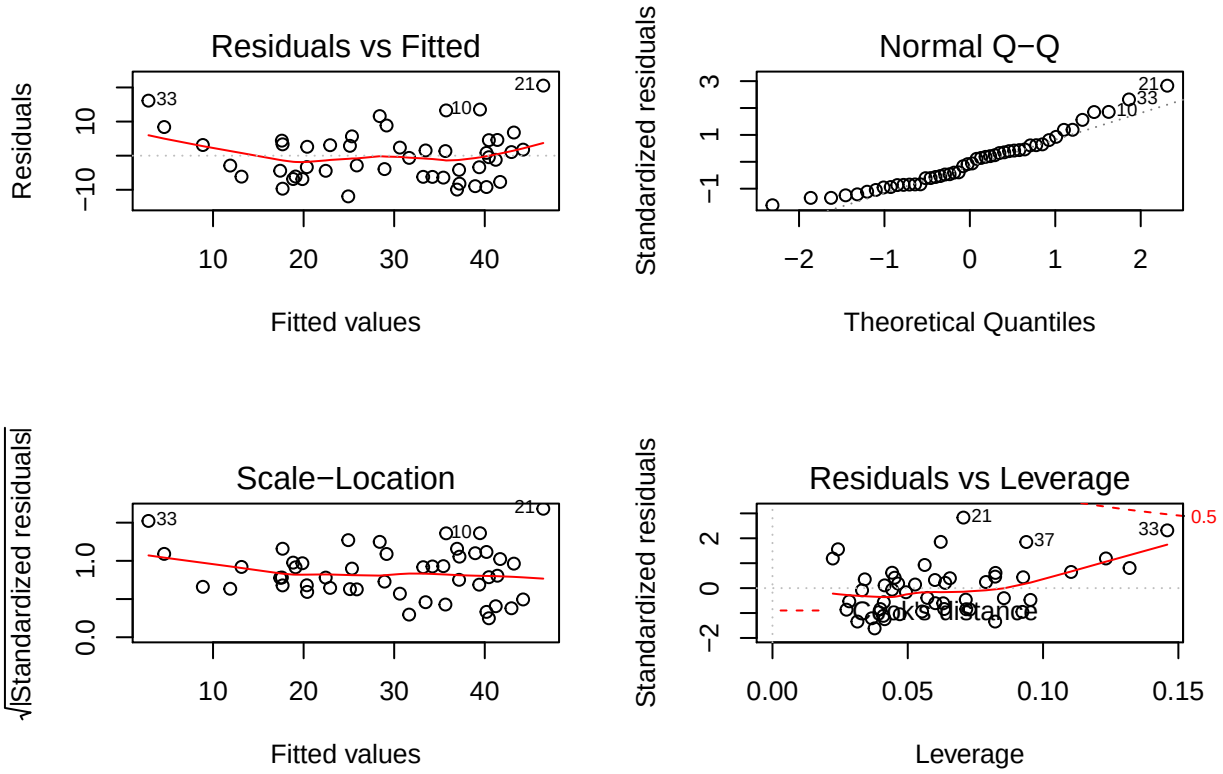
where y denotes **alumni giving rate**, x_1 denotes **graduation rate**, x_2 denotes **student&faculty ratio**. Apart from the previous method, We can use **stepAIC** method to directly construct the best multi-linear regression model from a set of candidate variables.The experiment result indicates that, AIC value becomes smaller when the variable **"Number of classes under 20"** is deleted from the regression model. The lowest AIC value is **196.7**. Since smaller AIC value means better fitting effect, we should construct a regression model with **"Graduation Rates"** and **"Student & Faculty Ratios"** as independent variables. The linear regression model fomula is constructed as

$$y = -19.1063 + 0.7557x_1 - 1.2460x_2 \tag{2}$$

The result given is the same as the estimated model.Both results suggest that the alumni giving rate is significantly affected by graduation rates and student & Faculty Ratios, instead of number of classes under 20.

1.4.Check the model assumptions

In Q-Q graph, points closely cluster around the line, which proves that the assumption of **normality** is satisfied; there is no reason to assume that graduation ratio and faculty to student ratio is related. Therefore, the assumption of **independence** is satisfied;from graph one, we can observe that residuals has no systematic relationship between residuals and the predicted values. The model well captures systematic variance in the data, thereby proves that the assumption of **Linearity** is satisfied; the Scale-Location graph shows that the points form a random band around the horizontal line. Hence the assumption of **Homoscedasticity** is satisfied.



Part 2

2.1. Calculate the mean of Fertility and partition the provinces into two groups

The mean of fertility in stated provinces is **70.14255**. Experiment details are omitted here.

2.2. Use logistic regression to show the relationship between y and the other variables and then interpret the regression results

The experiment result shows that **Agricluture**, **Examination** and **Education** are negatively related to Fertility, with remaining factors positively related to Fertility. P-values for **Examination** and **Agriculture** are 0.0203 and 0.0165 respectively. While p-values for other factors are all beyond the range of significance. The experiment result indicates that fertility in selected provinces is significantly related to **Agriculture** and **Examination** under significance level of 0.05. It implies that fetility is closely related to agriculture situation and examination circumstance in the provinces.

2.3. Choose a model selection criterion, for instances, AIC, BIC, adjusted R square or Cp, and use it to select a reasonable model

We can construct the best multi-linear regression model from a set of candidate variables. Under AIC criterion, regression model with smaller AIC value is considered better. The experiment result indicates that AIC value gets smallest(at -85.27) when the variables "Catholic" and "Education" are deleted from the regression model. Hence we should construct a regression model with **"Infant.Mortality"**, **"Agriculture"** and **"Examination"** as independent variables. The regression model constructed is:

$$y = 1.05056 - 0.00811x_1 - 0.05017x_2 + 0.03501x_3 \tag{3}$$

where x_1, x_2, and x_3 denotes **Agriculture**, **Examination**, and **Infant Mortality** respectively.