Data Mining

Assignment 1

Vivek Pandit 130010009

Q1. There is a lot of missing data i.e many labels have been marked NA in the csv file.

Q2. The strategy for imputing the missing data is by fitting a least square line on the remaining data, i.e after removing the incomplete data. Then using this best fit line, the missing labels for the given x coordinates are given by the value of the line at that value of x.

Q4. Least squares fit would be used to fit the model to predict y given x

Q5. The parameters are

Coefficients:

(Intercept) tab$x

6.083 2.232

Q6. Yes, the model accurately justifies the data because the summary of the best fit line gives

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.08283 0.35647 17.06 <2e-16 \*\*\*

tab$x 2.23236 0.04416 50.55 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.662 on 98 degrees of freedom

Multiple R-squared: 0.9631, Adjusted R-squared: 0.9627

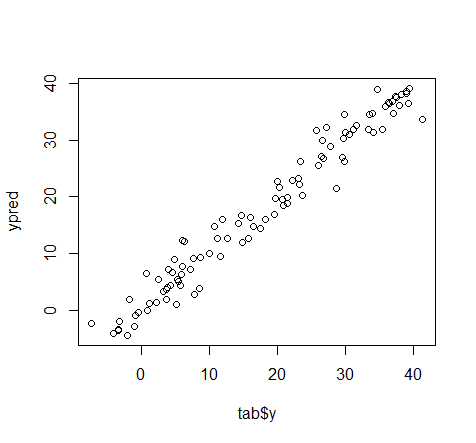
F-statistic: 2555 on 1 and 98 DF, p-value: < 2.2e-16

Since the value of R-squared and adjusted R squares is very close to 1 and the value of p for every coefficient is less than 0.05, hence this accurately represents the data.

Q7. RMSE = 2.635021

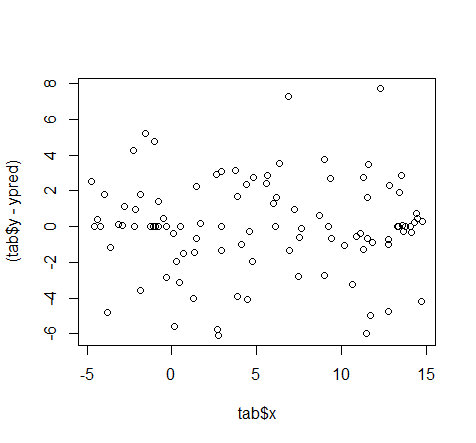
MAE = 1.883715

Q8.



This is a plot of Y\_predicted vs Y. Ideally, it should be a straight line inclined at 45 degree. The above plot strongly resembles the ideal case and shows a strong correlation between the two.

Q9.



The above plot is that of the residuals vs the value of x. The above plot indicates that there is high frequency at zero and as the distance from zero increases, the number of points decreases. Hence, this is representative of the normal distribution of the residuals i.e the actual model has been simulated using the error term as a normal distribution with mean zero.