STA303 Assignment 1

Joyce Nguyen

Feb 3rd, 2023

Question 2

The method for this question is implemented in R.

2.1. Estimate the coefficients of the regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

Using the provided algae growth model, the estimated coefficients of the regression model are:

Coefficients	Notation	Estimate
Intercept	β_0	0.4096
Days	β_1	0.0005
Dosage of silver metal (mg)	β_2	-0.0756
Days:Dosage of silver metal (mg)	β_3	0.0003

The resulting estimate regression model is:

$$\hat{y}_i = 0.4096 + 0.0005 x_{1i} - 0.0756 x_{2i} + 0.0003 x_{1i} x_{2i} + \epsilon_i$$

2.2. Test the goodness of the overall model

Looking at the values from the summary of linear regression model, we can test its goodness:

- The median value of residuals is -0.0019 which is approximately equal to 0, telling us that our residuals are somewhat symmetrical (normally distributed) and that our model is predicting evenly at both the high and low ends of the dataset.
- Residual standard error of 0.0575 tells us that the regression model predicts the growth of algae with an average error of about 0.0575. This number is sufficiently small to say that the model's prediction line is very close to the actual values, on average.
- Adjusted R^2 is equal to 0.743, saying that 74.3% of variance in the growth of algae (dependent variable) is explained by our independent variables.
- The value of F-statistic is very large (54.02) and the p-value is significantly small ($5.383 * 10^{-16}$), showing that there is sufficient evidence to reject the null hypothesis that there is not relationship between dependent and independent variables.

To sum up, we conclude that there is strong evidence that a relationship does exist between algae growth versus one of the predictors (days, dosage of silver metal, and the interaction term between days and silver metal dosage); and the linear model that we fit is good to use.

2.3. Hypotheses Testing

In this part, we wonder whether the the interaction term between days and dosage of silver metal is really affecting the growth of algae, or it is sufficient to be removed. Thus, we consider the hypotheses:

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

Now wee want to breakdown the sum of squares to see if the model does a good job at explaning the trend. This can be done with a global ANOVA test. The test statistic that we will be looking at is the F-statistic, which is calculated by the formula:

$$F = \frac{SS_{reg}/p}{RSS/(n-p-1)} \text{ where } F(p, n-p-1)$$

The F-statistic we get for the interaction term is 0.1074, which is very small based on the probability distribution. Further more, the p-value = 0.7444 for this test very large. Thus, we say that the null hypothesis is true and that we would want to change the model by removing the interaction term.

To confirm our conclusion, we use a T-test on each individual coefficient to test whether the interaction term has a linear relationship with the algae growth in the presence of other predictors. We see that the resulting p-value for interaction term of 0.744 is much higher than the standard level of 0.05. This means we fail to reject the null hypothesis, indicating that the interaction term does not significantly affect the growth of algae (dependent variable). So even if we drop the interaction term out of the model, the predicting ability of the model is not strongly influenced.

The new model that we adopt is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

Furthermore, even after dropping the interaction term we ended up with an adjusted R^2 equal to 74.7%, indicating that barely any information was lost by removing this predictor. Hence, this model looks good to go.

2.4. Lack of fit test

Based on the statistics in the previous part, we have decided to removed the interaction term from the model. But we are not sure how well the reduced model fits well compared to the full one. So now we move forward to performing a lack of fit test, or a partial F test to consider how the conditional relationship between days and dosage of silver metal (our predictors) and the algae growth (our response) changes if the interaction term is removed.

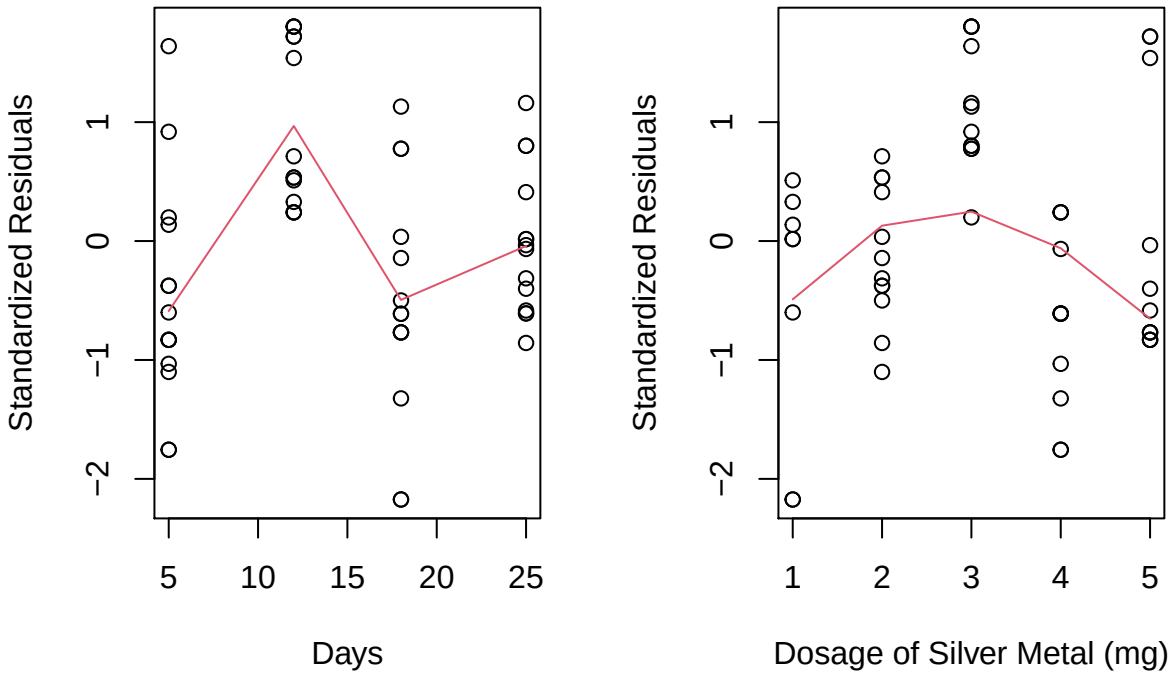
The F-statistic for comparing our reduced and full models based on ANOVA is given by:

$$F = \frac{(RSS_{reduced} - RSS_{full})/(df_{reduced} - df_{full})}{RSS_{full}/df_{full}}$$

The p-value of ANOVA test equals 0.7444, which is very high compared to the threshold of 0.05. Thus, we cannot adopt the full model and we conclude that the reduced is the final model.

2.5. Plot standardized residuals vs. predictors

For the new reduced model, we would like to plot standardized residuals against each predictor variable to check for the assumption of linearity.



- For the plot of standardized residuals versus predictor days, observations seem to scatter quite randomly around the residual equal to 0 line. However, we can also observe some points at around the 12th day are higher from the basic pattern of other residuals. It indicates that the linearity is not perfectly valid.
- The second plot of dosage of silver metal highlights a quite concerning quadratic pattern. Although the points are around the residual equal to 0 line, it is not randomly scattered. It looks sufficient to say that linearity assumption is not valid here.

2.6. VIF and multicollinearity

Sometimes for a multiple regression model, there might be a strong correlation between the independent variables. This is called the problem of multicollinearity and it will reduce the precision of the estimated coefficients, which weakens the statistical power of the regression model. Hence, it is necessary to test the multicollinearity problem of our model.

We will rigorously test it using the Variance Inflation Factor (VIF), which is the term $\frac{1}{1-R_j^2}$ from the variance formula

$$Var(\hat{\beta}_j = \frac{1}{1-R_j^2} * \frac{\sigma^2}{(n-1)S_{x_j}^2} \text{ where j=1,...,p}$$

After calculating in R, we ended up with VIF = 1 for both predictors (days and dosage of silver metal), which is an indicative of no multicollinearity. We conclude that there exists no multicollinearity in our final model.

Question 3

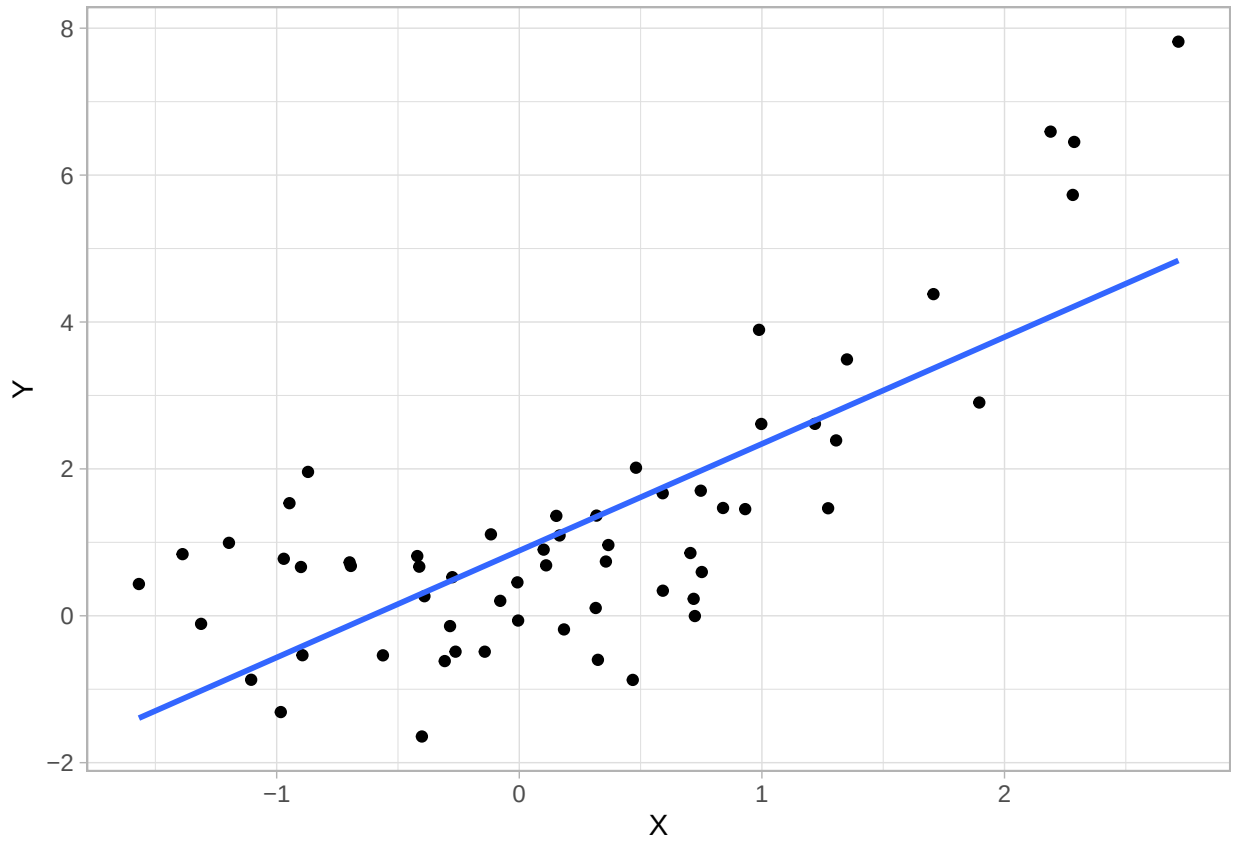
The method for this question is implemented in R.

In this question, 60 observations were simulated from the following polynomial regression:

$$Y = 0.5X + X^2 + \epsilon, \epsilon \in N(0, 1)$$

3.1. Simple linear regression between X and Y

For this part, we are curios about the linear regression performance between X and Y, so we plot a simple linear regression between these two:



Looking at the plot, there seems to exist a linear relationship between X and Y. The trend is shown even more clearly by looking at the fitted line where the observations are scattered closely to. Hence, the simple linear regression might be a candidate model (reduced model) that can fit the data. We will save this aside and compare it to the full model later on.

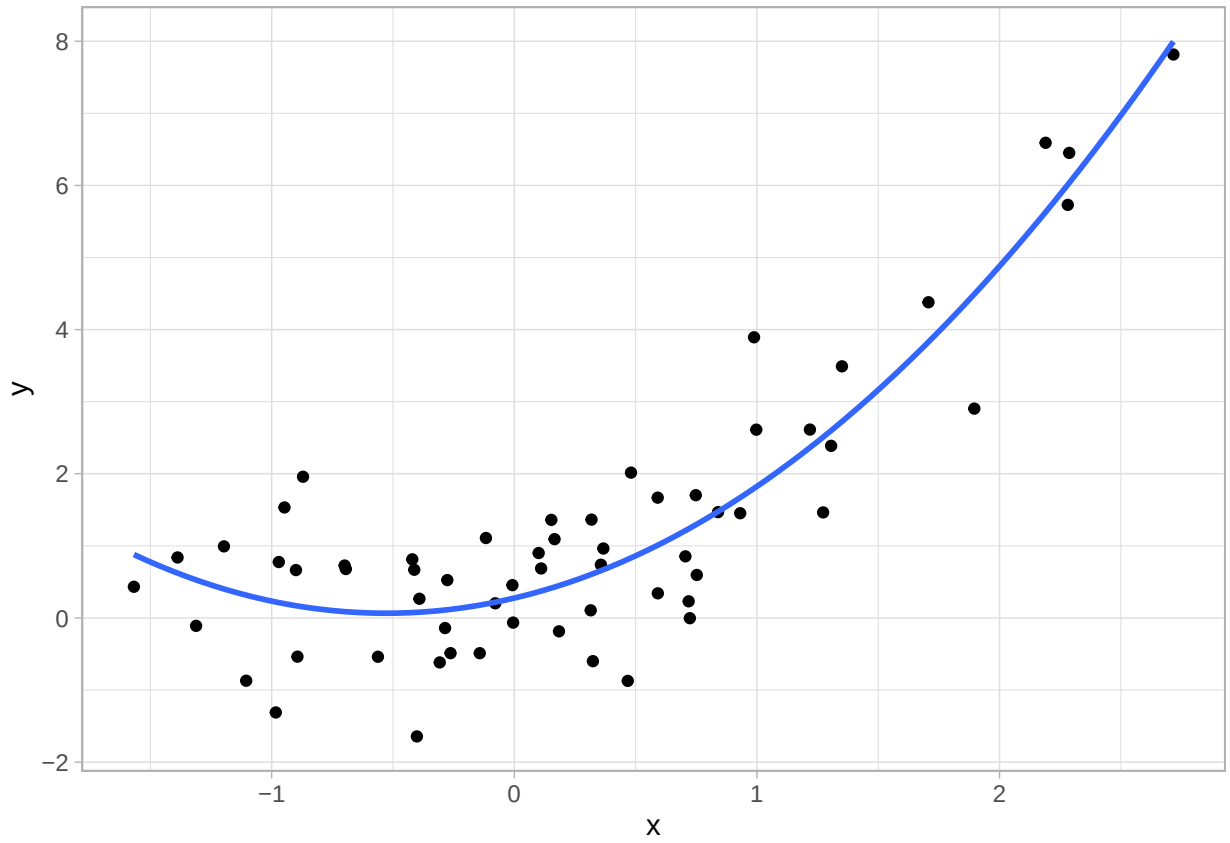
3.2. Estimate the coefficients of the full model

Using the stimulated model, the estimated coefficients of the correct polynomial regression model (full model) are:

Coefficients	Notation	Estimate
Intercept	β_0	0.2754
X	β_1	0.7957
X^2	β_2	0.7531

3.3. Plot of X versus Y (full model)

In order to better understand the performance of the full model, here we plot X versus Y with a smooth polynomial curve for this model:



3.4. Use F-test from ANOVA to compare models

As we are having two candidate models: the simple regression (reduced) model and the polynomial (full) model. It is time to decide which one should be used. We will conduct a partial F-test to test whether there is a statistically significant difference between two models (alternative hypothesis) or the full model can be replaced by the reduced one as the difference is not very considerable (null hypothesis).

The F-statistic for comparing our reduced and full models based on ANOVA is given by:

$$F = \frac{(RSS_{reduced} - RSS_{full})/(df_{reduced} - df_{full})}{RSS_{full}/df_{full}}$$

By conducting the test in R, we ended up with the p-value from F-test equals to $3.763 * 10^{-11}$, which is much smaller than the significant level of 0.05. So there is sufficient evidence to reject the null hypothesis that all removed predictors were not necessary, and the test suggests to keep the correct polynomial model (full model).

3.5. Coefficient of determination of final model

Since the full model is our final model, we should take into account how good the model is at predicting. So we look at the model's coefficient of determination $R^2 = \frac{RSS}{TSS}$, where RSS is the sum of squares of residuals and TSS is the total sum of squares.

The adjusted R^2 we got is equal to 0.79, which means 79% of the variation can be explained by the model. This model is sufficiently good at predicting outcomes. However, we have to keep in mind the fact that there are still 21% of the outcomes that the model cannot take into account of.

3.6. Leverage/Outlier points

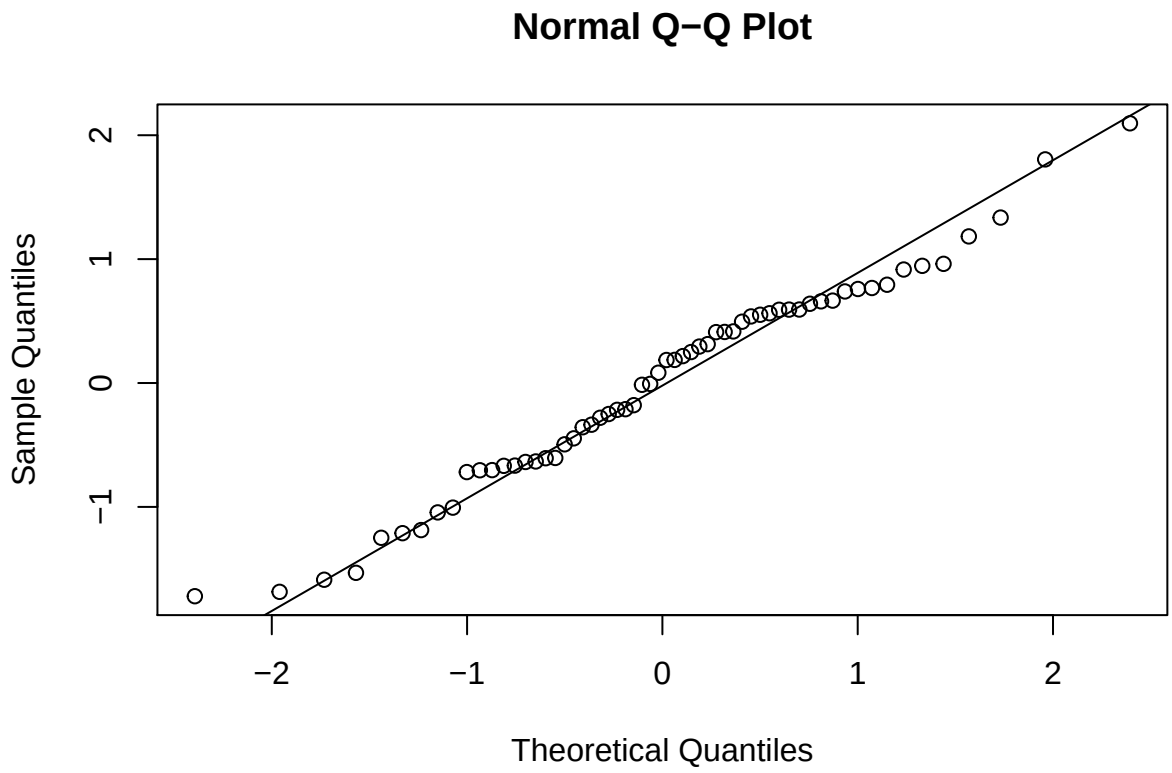
Now we move to another step of checking whether our model is doing a reasonable job at fitting the data by identifying problematic observations. There are two types of points we have to identify that might influence the goodness of our model:

- Outliers: points that do not fall on or near the trend of other observations, i.e. the regression line.
- Leverage points: points that lie very faw aware from the horizontal mean (mean of predictor).

3.6-a. Identifying leverage/outlier points that might potentially cause an issue

To begin with, we look at the normal QQ plot to check for normality assumption, as well as to have an overview of any outliers or leverage points in the dataset.

```
fsim_res <- full_sim$residuals
qqnorm(fsim_res)
qqline(fsim_res)
```



At first glance, the data is quite normally distributed. However, we can identify some suspected outliers and leverage points at the head and tail of the dataset. Now we will take a closer look at each type of problematic points.

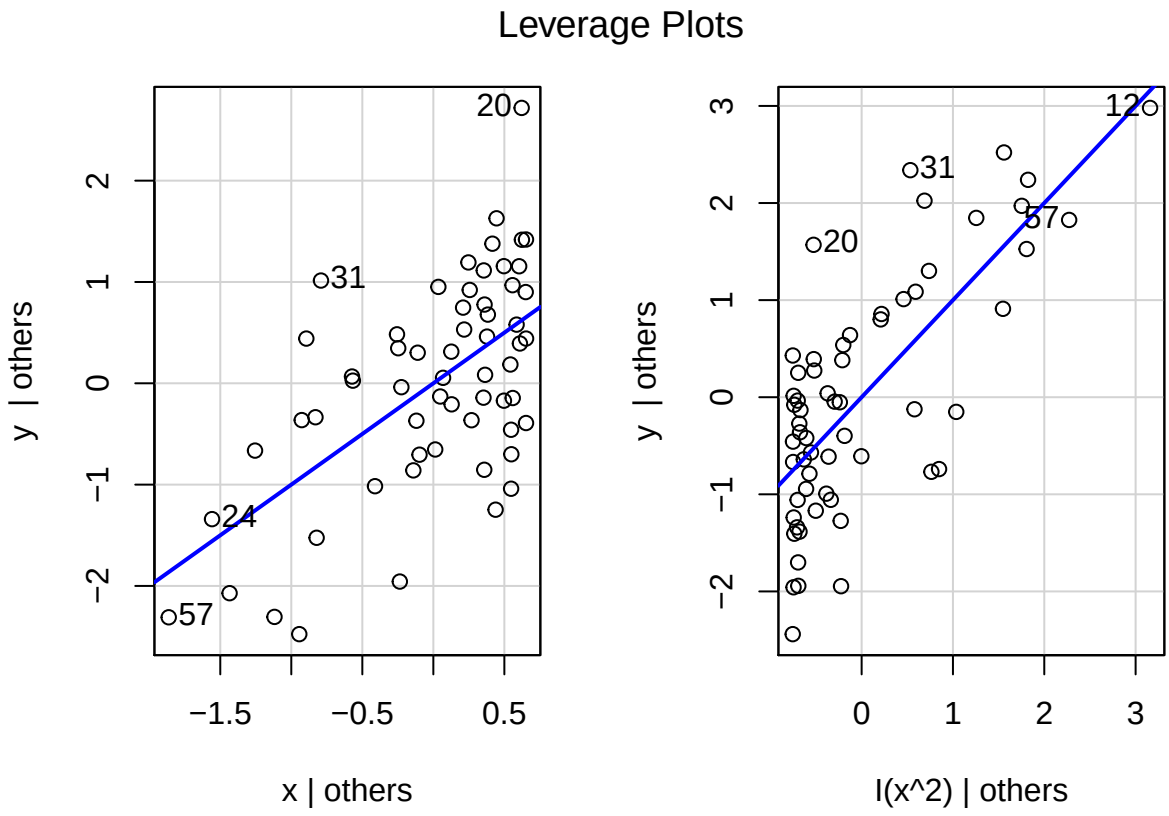
For outliers, we test the hypothesis:

$$H_0 : \text{Data has no outliers}$$

$$H_1 : \text{Data has at least one outlier}$$

The resulted p-value from the outlier test we performed in R is 0.012, suggesting that we have some outliers in the model for the growth of algae.

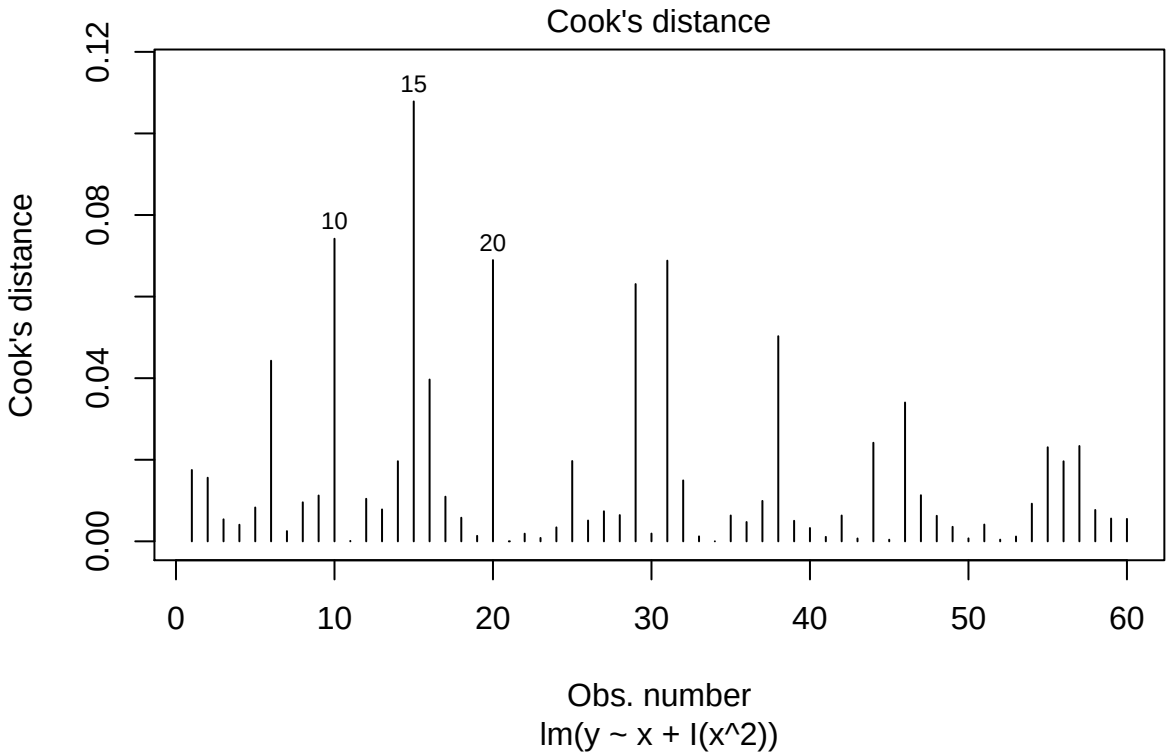
For leverage points, we first look at some plots for each predictor:



From the plots of each predictor against response, we see that there are some leverage points as expected. We also conducted the hat matrix method to determine points of high leverage that can potentially cause issues with the model, and the result shows that obersvations number 1, 10, 12, 13, 24, 44, and 57 might induce some issues.

3.6-b. Identifying influential observations

To get a clearer view of which observation has the highest influence on the our model, we look at the plot of Cook's distance below:



The Cook's distance shows that observations number 10, 15, and 20 have the highest influence on the entire regression line.

While the Cook's Distance looks at the effect of a single observation on all fitted values, we can also conduct DFFITS test to look at that effect of that observation on its own fitted value. With that saying, the DFFITS test tells us that observations number 10, 15, 20, and 31 are influential.

After taking a look at both test, we see some overlapping between the results they give. Hence, it is sufficient to suggest removing influential points number 10, 15, and 20 to guarantee the precision of the predicting ability of our model.