BTRY 6020 Homework II

NAME: ANDRES CASTANO

NETID: AC986

DUE DATE: 8:40 am Monday February 13

Question 1

As an alternative to dangerous insecticides, a chemist is working on a synthetic pheromone (a type of hormone involved in mating behavior) to be used as a bait to attract destructive insects into traps. Six different levels of the hormone are used in this study: (10, 20, 30, 30, 50, and 60). There are 60 traps and 10 traps are randomly assigned to each of the six doses of the hormone.

Data on the number of insects caught per trap and dose appear in the Excel file Hwk1Q1DatSp17.xlsx.

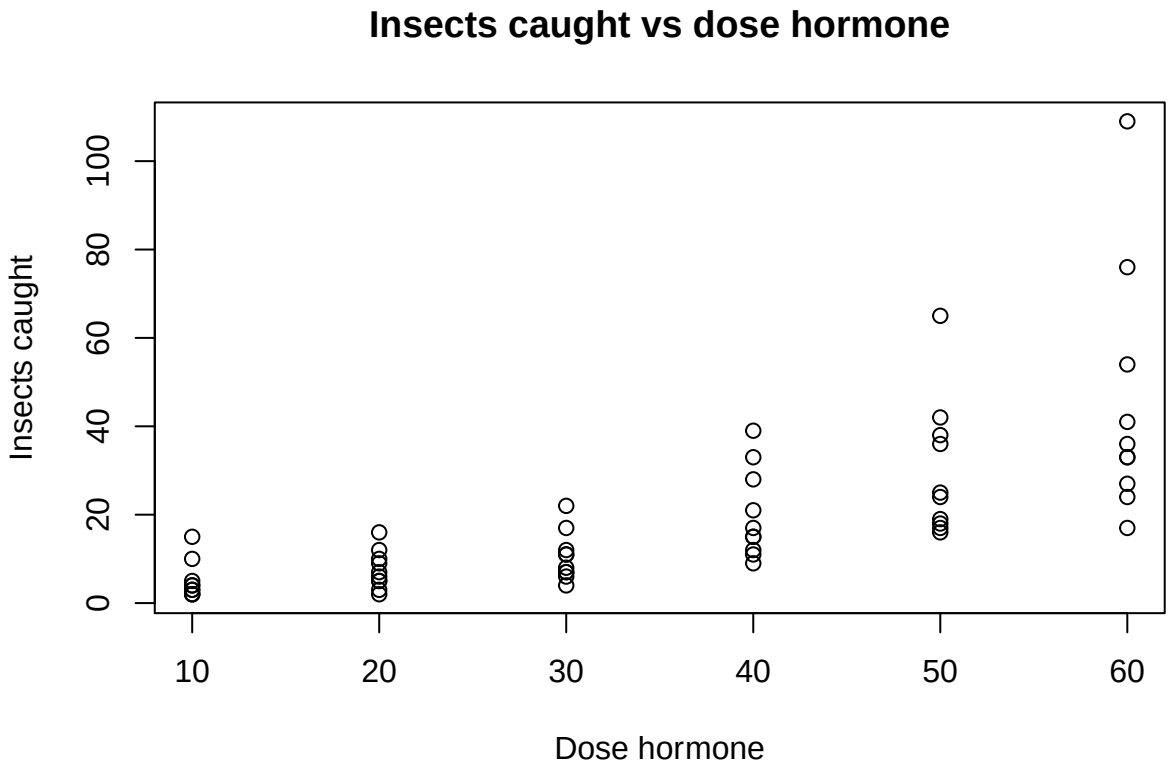
I have decided to change the name of the variables in the excel file to simplify the analysis in R studio. Y = "Insects_caught" and X = "Dose_hormone".

```
library(readxl)
data_insec = read_excel("Hwk2Q1DatSp17.xlsx")
names(data_insec)
```

```
## [1] "ObsNumber"      "Dose_hormone"   "Insects_caught"
```

a) Plot the number of insects caught against the dose of the hormone and assess curvature in the data (ALWAYS REMEMBER to first plot your data in a regression analysis).

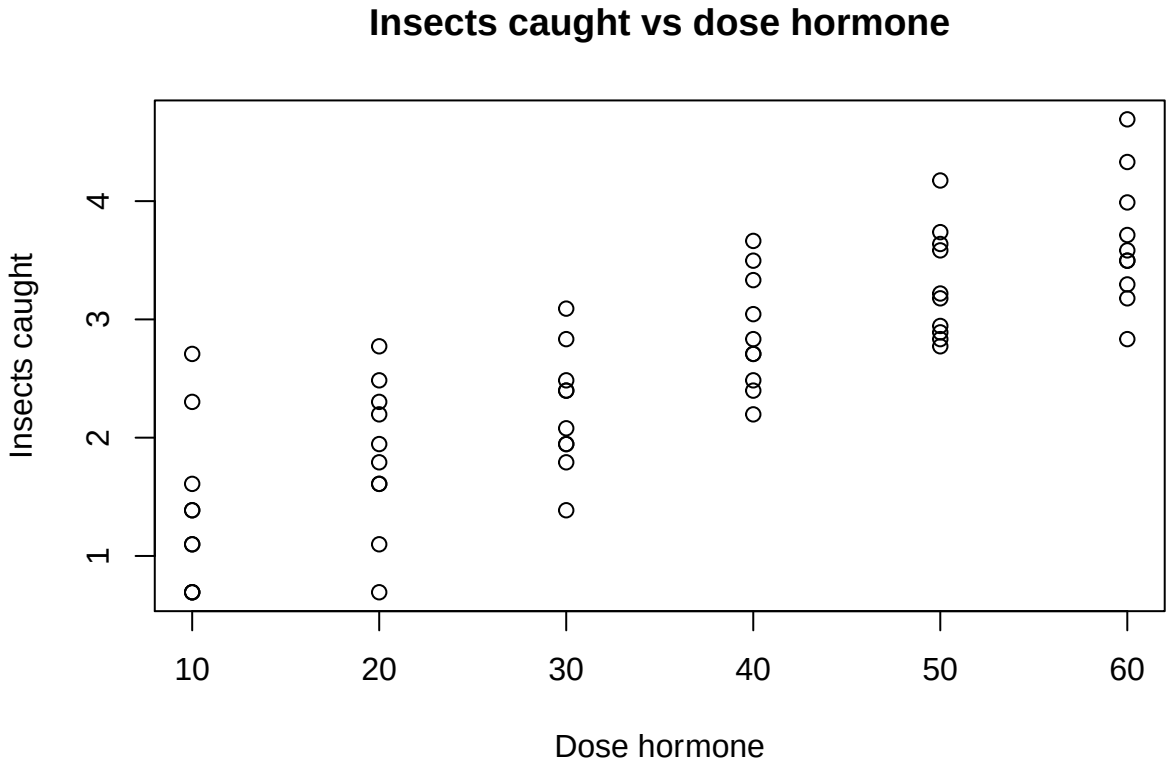
```
plot(data_insec$Dose_hormone, data_insec$Insects_caught , xlab = "Dose hormone", ylab = "Insects caught"
```



The graph depict that the relation between insect caught per trap and dose is not linear. The insects caught seems to be increasing at an increasing rate with the dose of hormone (exponential). To apply the SLR model, we need this relationship to be linear.

b) Plot the natural log of insects caught against the dose of the hormone and assess curvature in this relationship.

```
plot(data_insec$Dose_hormone, log(data_insec$Insects_caught), xlab = "Dose hormone", ylab = "Insects ca
```



After transforming the dependent variable (insects caught) to logarithm (natural logarithm), we can see more clearly that the relationship between the variables is more linear. The natural logarithm transformation of Y cause that Y goes down, which in turn help to linearize the relantionship between X and Y. In other words, help to stabilized the variance.

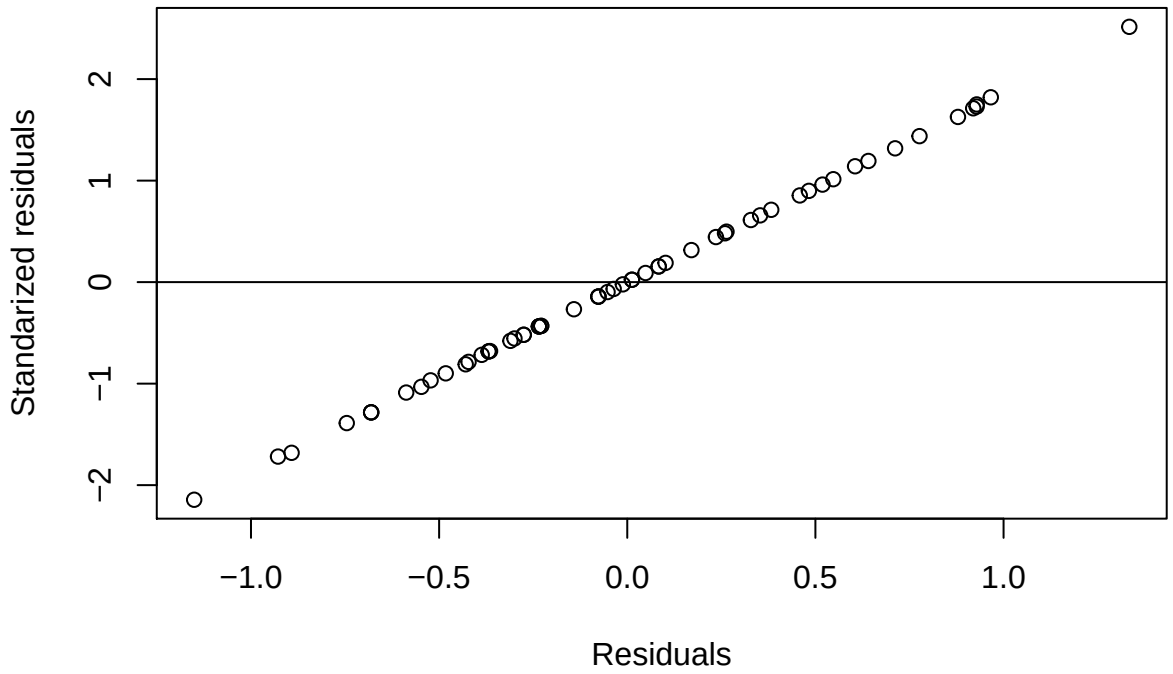
c) Regress the natural log of insects caught on the dose of the hormone obtaining residuals, standardized residuals, predicted values, leverage, and Cook's distance and answer the following questions: (R function: `lm()` for fitting linear model)

```
lm1.lm = lm(log(Insects_caught)~Dose_hormone, data = data_insec)
summary(lm1.lm)
```

```
##
## Call:
## lm(formula = log(Insects_caught) ~ Dose_hormone, data = data_insec)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15099 -0.36852 -0.06452  0.36046  1.33421
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.90355    0.16044   5.632 5.47e-07 ***
## Dose_hormone  0.04703    0.00412  11.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.545 on 58 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.6867
## F-statistic: 130.3 on 1 and 58 DF,  p-value: < 2.2e-16
```

i) Plot the residuals against the standardized residuals. What does this plot reveal? (R hint: `rstandard()` for getting standardized residuals.)

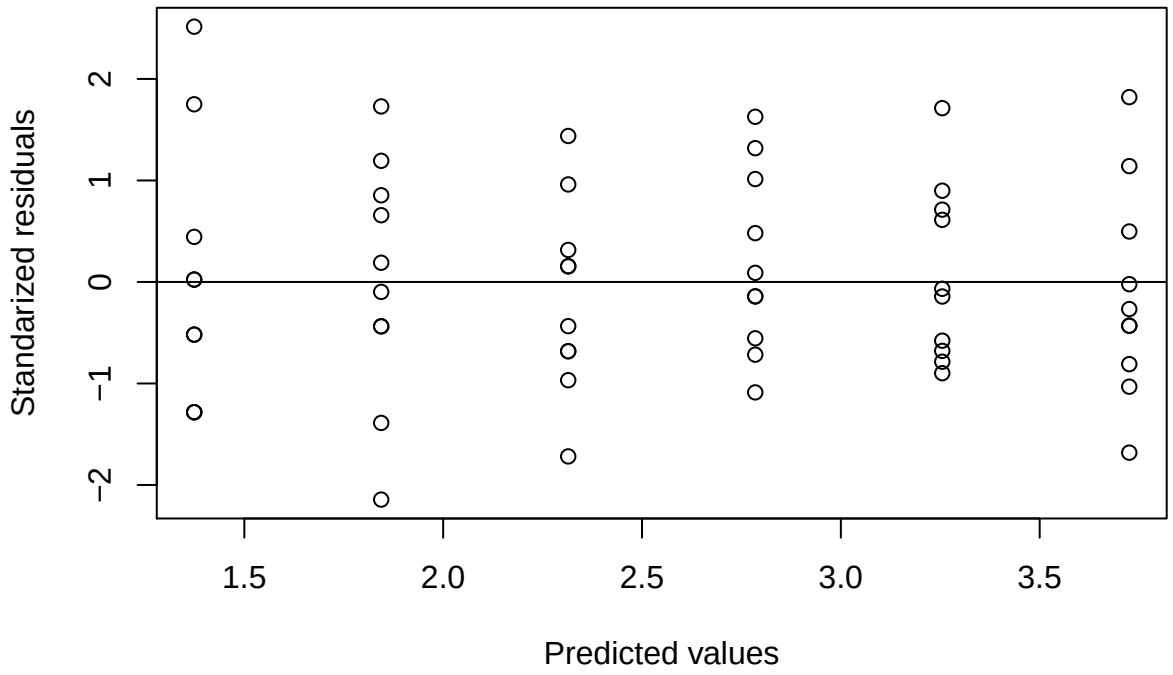
```
lm1.stdres=rstandard(lm1.lm )
plot(lm1.lm$residuals, lm1.stdres, ylab="Standarized residuals", xlab="Residuals", abline(0,0))
```



This plot is a good sign that we do not have outliers in our data. All the residuals are between -3 and +3 standard deviations from the mean. We also notice only two of these standardized residuals lie above +2 or below -2, which is about what we would expect with this number of observations. And none go much beyond these bounds, implying an absence of outliers.

ii) Plot the standardized residuals against the predicted values. Assess whether the assumption of equal variance is valid or not.

```
plot(lm1.lm$fitted.values, lm1.stdres, ylab="Standarized residuals", xlab="Predicted values", abline(0,
```



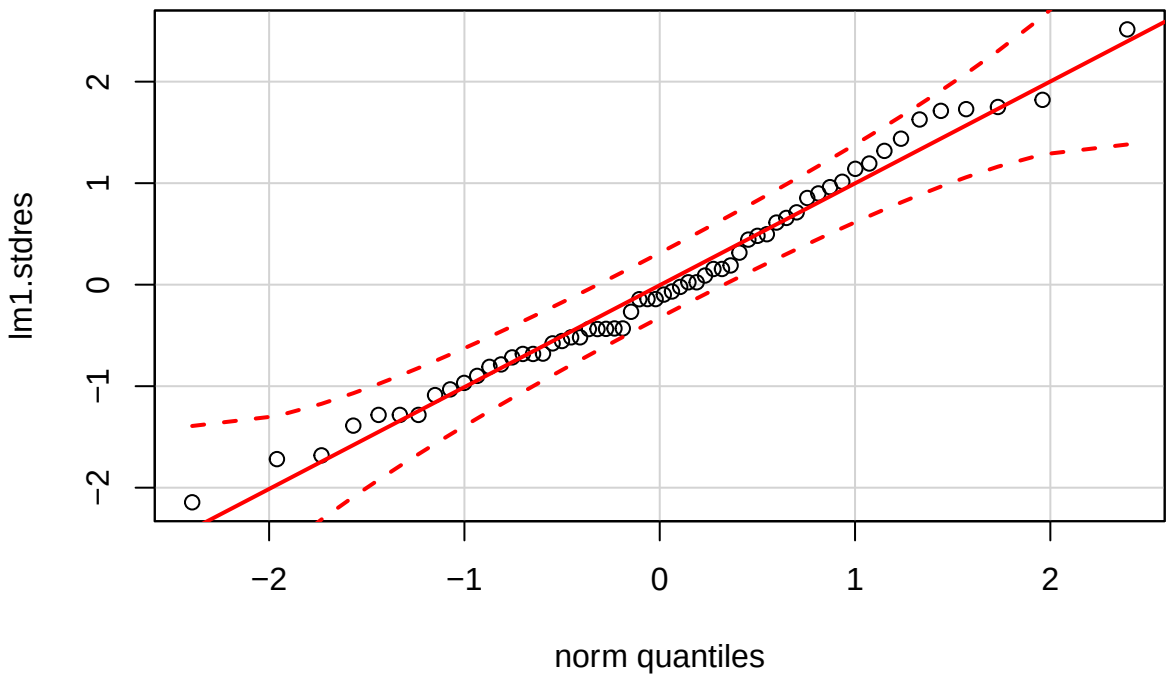
The standardized residuals are distributed evenly above and below the zero line across different predicted values. So the equal variance assumption of the residuals seems valid here.

iii) Assess the normality of the standardized residuals.

We can assess the normality of the standarized residuals in two ways: 1) make a histogram or 2) make a normal quantile comparison plot. We are going to try both:

```
hist(lm1.stdres)
library(car)
```

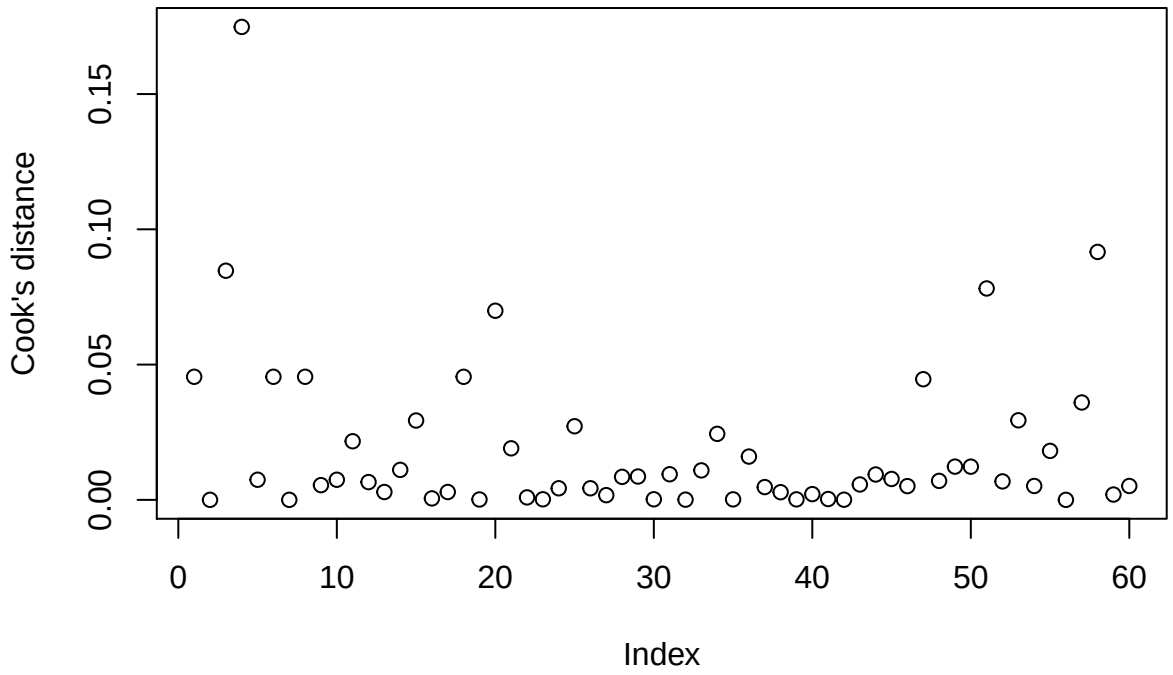

```
qqPlot(lm1.stdres)
```



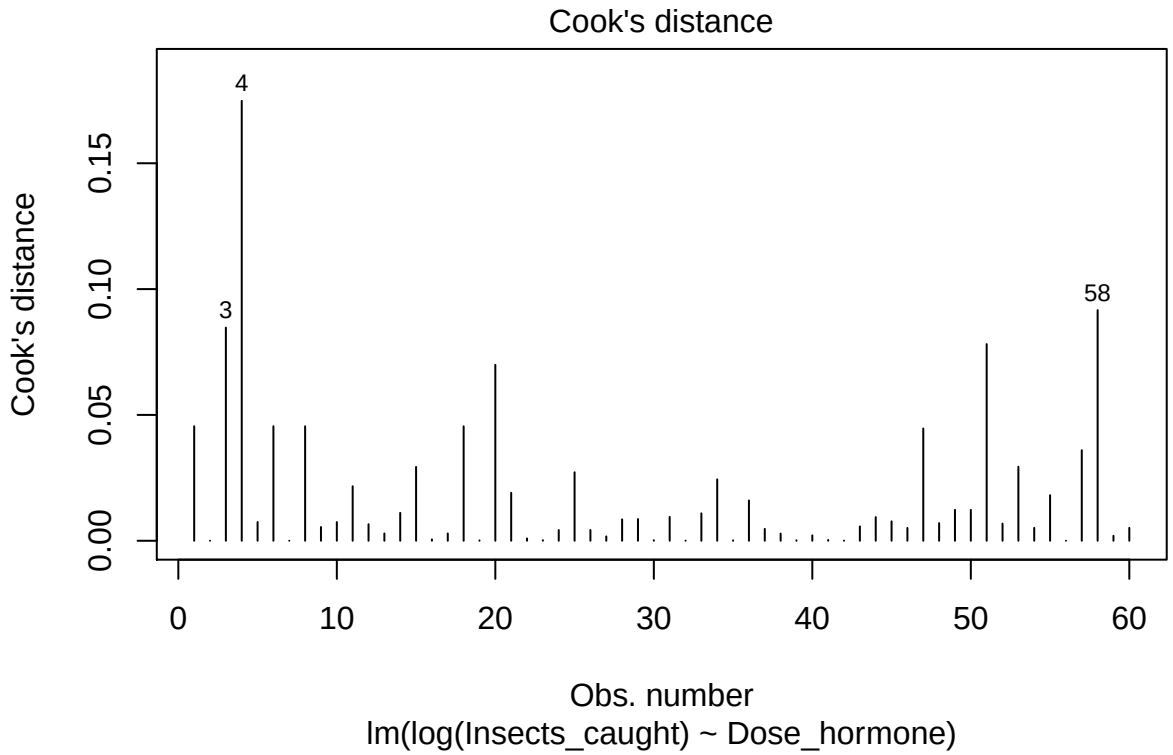
The histogram as well as the qqplot depict that the standarized residuals are roughly normal.

iv) Plot leverage and Cook's distance against observation number. Are there any data points with unusually high leverage? Are there any influential data points? (R hint: `hat(model.matrix())` generates leverage for all points; `cooks.distance()` generates Cook's distance.)

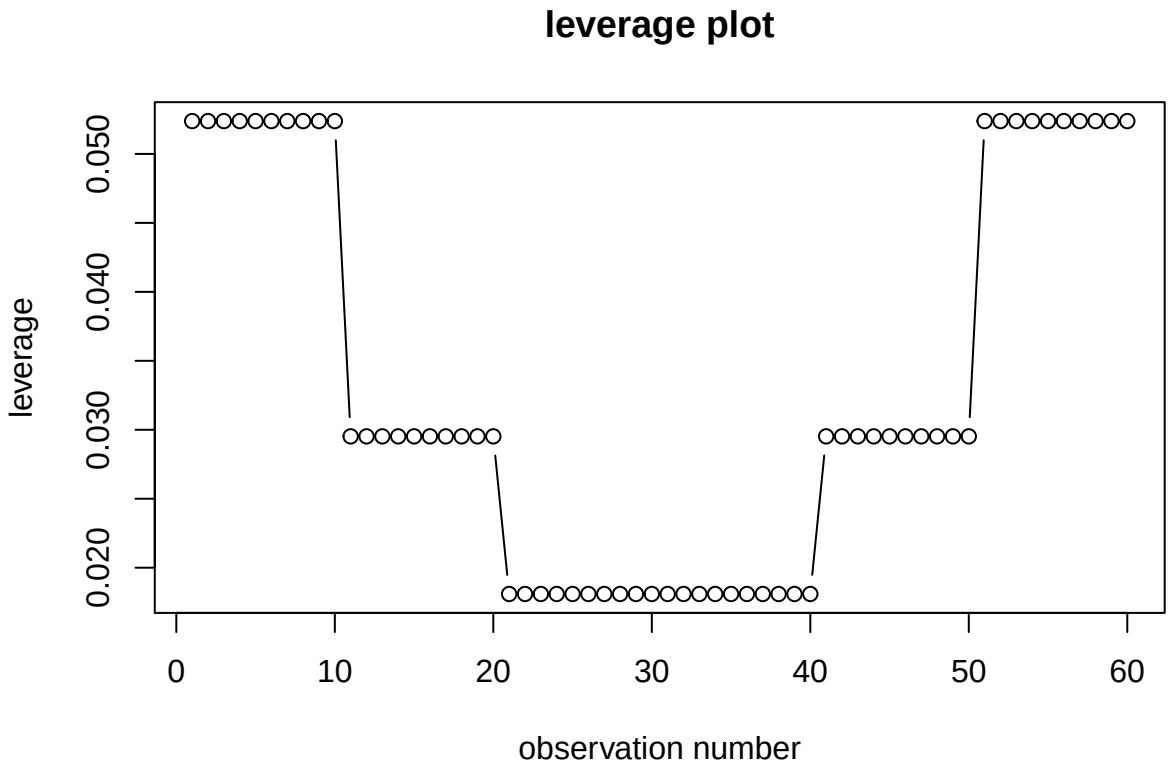
```r
# cook's distance
cookd = cooks.distance(lm1.lm)
plot(cookd, ylab = "Cook's distance")
```




```r
# another way to calculate the cook's distance (preferable)
plot(lm1.lm, which = 4)
```



```r
# Leverage plot
leverage = hat(model.matrix(lm1.lm))
plot(leverage, type="b", xlab="observation number", ylab="leverage", main="leverage plot")
```

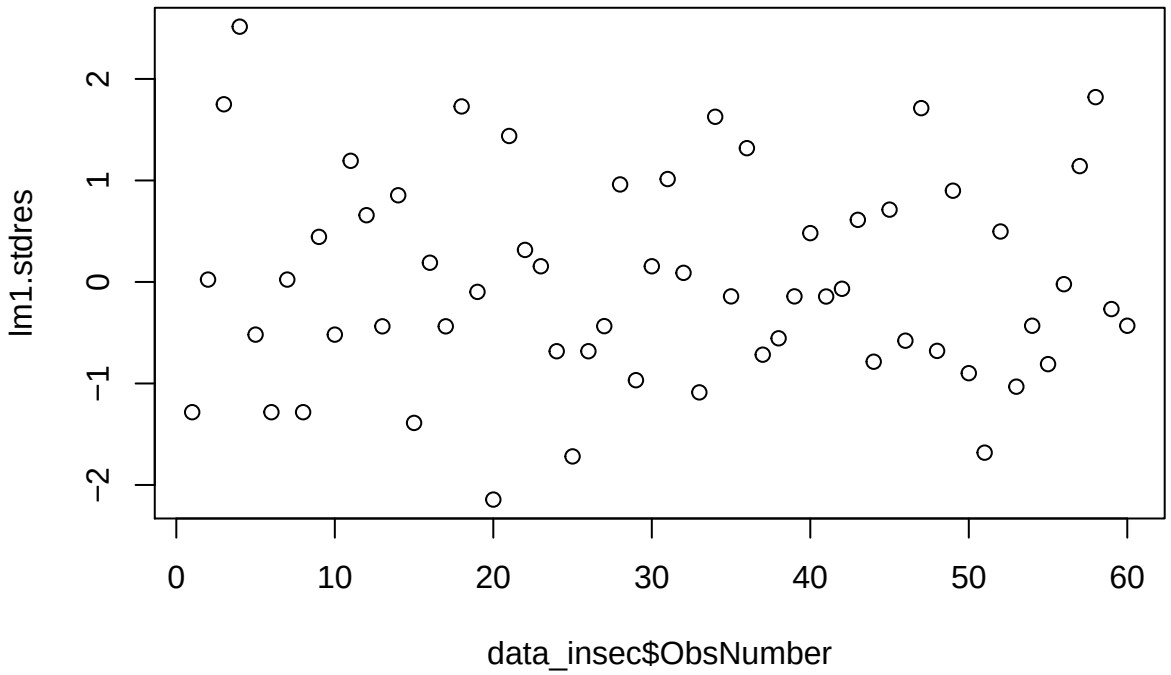



```r
mean(leverage)
```

```
## [1] 0.03333333
```

Since all the leverage are smaller than $\frac{2p}{n} = \frac{4}{60} = 0.067$, there are no high leverage points. Futhermore, all the values of the cook's distance are less than 1, which means that we do not have a single data point that highly influence the estimated parameters. However, there is a point that despite that is less than one is certainly more influential than the others (observation number 4). In the graph below, we plot the standarized residuals (X) against the index of data (X). We can observe that observation number 4 is less than 3 standard deviations from the mean of the residuals, then it could be a possible value for a variable normally distributed.

```r
plot(data_insec$ObsNumber, lm1.stdres)
```



On the other hand, the average hat value is $\bar{h} = (k+1)/n = (2/60) = 0.03333$ (where k is the number of regressors excluding the constant and n is the number ob observations). We have many points with a individual leverage above the mean, however the difference is not much to consider those highly leverage points.

d) Compute a 95% confidence interval for the slope of the regression line and carefully state your conclusions. (R hint: `summary()` generates summary table for linear model.)

```
summary(lm1.lm)
```

```
##
## Call:
## lm(formula = log(Insects_caught) ~ Dose_hormone, data = data_insec)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15099 -0.36852 -0.06452  0.36046  1.33421
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.90355    0.16044   5.632 5.47e-07 ***
## Dose_hormone  0.04703    0.00412  11.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.545 on 58 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.6867
## F-statistic: 130.3 on 1 and 58 DF,  p-value: < 2.2e-16
```

A 95% confidence interval for the slope of the regression line is:

$$\hat{\beta}_1 \pm (t_{\frac{\alpha}{2},n-k}) * SE(\hat{\beta}_1)$$

Where:

$$SE(\hat{\beta}_1) = \sqrt{\frac{S_e^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$S_e^2 = \frac{\sum_i^n e_i^2}{n-k}$$

With k= number of parameters; n= number of observations; and $e_i = (Y_i - \hat{Y}_i)$

For our data, we are going to calculate the intervals as follows (based on the regression results):

```
n=60
k=2
B1 = 0.04703
SE_B1 = 0.00412
t_crit = -qt(0.025, df=58)
t_crit
```

```
## [1] 2.001717
```

```
lower = (B1) - (t_crit)*(SE_B1)
lower
```

```
## [1] 0.03878292
```

```
upper = (B1) + (t_crit)*(SE_B1)
upper
```

```
## [1] 0.05527708
```

```
# confidence interval is
c(lower, upper)
```

```
## [1] 0.03878292 0.05527708
```

```
# with the following command we can confirm our results
confint(lm1.lm)
```

```
##                    2.5 %     97.5 %
## (Intercept)  0.58239590 1.22471043
## Dose_hormone 0.03878255 0.05527567
```

Then, our 95% confidence interval is (0.039,0.055). Which means that we are 95% confident that our population slope parameter is between 0.039 and 0.055. In other words, We are 95% confident that the natural log of number of insects trapped increases between 0.0388 and 0.0553 for each unit increase in hormonal level.

e) Test to see if the number of insects caught increase with dose. State Hypotheses, Test Statistic, p-value, and conclusions.

Here we are interested in determine wheter $\beta_1 > 0$, then we can formulate our hypothesis as follows:

$H_0 : \beta_1 = 0$

$H_A : \beta_1 > 0$

Our test statistics is:

$test = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.04703}{0.00412} = 11.41$

In R, we can compute the p-value associate to this statistic as follows:

```r
t=(B1-0)/(SE_B1)
t
```

```
## [1] 11.41505
```

```r
p_value= pt(t,n-k, lower.tail = FALSE) # pvalue related with one side test, we are only interested if b
p_value
```

```
## [1] 9.17569e-17
```

At significance level of 0.05, we have evidence to reject the null hypothesis (p-value<0.05). Then, we conclude that the number of insects caught increase with the dose of the hormone.

f) Calculate and interpret a 95% confidence interval and a 95% prediction interval for the number of insects caught if a hormone dose of 40 is used. (R hint: `predict()` with *interval* argument set to "confidence" or "prediction")

The 95% confidence interval for a data point is define as follows:

$$\hat{Y} \pm (t_{\frac{\alpha}{2},n-k}) * S_e \sqrt{\frac{1}{n} + \frac{X_0 - \bar{X}}{(n-1) * (S_x^2)}}$$

In our case this definitions is translated into R as:

```r
summary(lm1.lm) # to extract the values nedeed
```

```
##
## Call:
## lm(formula = log(Insects_caught) ~ Dose_hormone, data = data_insec)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15099 -0.36852 -0.06452  0.36046  1.33421
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.90355    0.16044   5.632 5.47e-07 ***
## Dose_hormone  0.04703    0.00412  11.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.545 on 58 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.6867
## F-statistic: 130.3 on 1 and 58 DF,  p-value: < 2.2e-16
```

```r
n=60
k=2
B1=0.04703
BO=0.90355
SE_residuals=0.545
Y_estimated_40 = (0.90355) + (0.04703)*40
Y_estimated_40
```

```
## [1] 2.78475
```

```r
mean_x=mean(data_insec$Dose_hormone)
var_x=(var(data_insec$Dose_hormone))
```

```
t = -qt(0.025, n-k)
se_Y_estimated_40 =  (SE_residuals) * sqrt(((1/n) + (40-mean_x)/((n-1) * var_x)))
ci_lower = Y_estimated_40 - (t) * (se_Y_estimated_40)
ci_upper = Y_estimated_40 + (t) * (se_Y_estimated_40)
# the confidence interval is
c(ci_lower, ci_upper)
```

```
## [1] 2.642709 2.926791
```

```
# In the original scale of the Y variable the 95% confidence interval is:
c(exp(ci_lower), exp(ci_upper))
```

```
## [1] 14.05121 18.66764
```

we can verify our calculations with the following line command. (The difference is quite small and it should be explained by our incomplete use of all the decimals. The results are in log scale).

```
conf_interval = predict(lm1.lm, newdata=data.frame(Dose_hormone=40), interval="confidence", level = 0.95
conf_interval
```

```
##        fit      lwr      upr
## 1 2.784718 2.637969 2.931466
```

```
exp(conf_interval)
```

```
##        fit      lwr      upr
## 1 16.19524 13.98477 18.75511
```

There is a little difference between our calculations and the calculations that the software made. Using the results in the original scale of Y, we can say that we are 95% confident that when the hormone dose is 40, the expected number of insects caught will be between 14.05 and 18.68.

On the other hand, the interval prediction for a data point is defined as:

$$PD(\hat{Y}) \pm (t_{\frac{\alpha}{2},n-k}) * S_e \sqrt{1 + \frac{1}{n} + \frac{X_0 - \bar{X}}{(n-1) * (S_x^2)}}$$

We can get this interval in R as follows:

```
summary(lm1.lm) # to extract the values nedeed
```

```
##
## Call:
## lm(formula = log(Insects_caught) ~ Dose_hormone, data = data_insec)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15099 -0.36852 -0.06452  0.36046  1.33421
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.90355    0.16044   5.632 5.47e-07 ***
## Dose_hormone  0.04703    0.00412  11.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.545 on 58 degrees of freedom
```

```
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.6867
## F-statistic: 130.3 on 1 and 58 DF,  p-value: < 2.2e-16
```

```
n=60
k=2
B1=0.04703
BO=0.90355
SE_residuals=0.545
Y_predict_40 = (0.90355) + (0.04703)*40
Y_predict_40
```

```
## [1] 2.78475
```

```
mean_x=mean(data_insec$Dose_hormone)
var_x=(var(data_insec$Dose_hormone))
t = -qt(0.025, n-k)
se_Y_predict_40 =  (SE_residuals) * sqrt((1 + (1/n) + (40-mean_x)/((n-1) * var_x)))
pi_lower = Y_predict_40 - (t) * (se_Y_predict_40)
pi_upper = Y_predict_40 + (t) * (se_Y_predict_40)
# the confidence interval is
c(pi_lower, pi_upper)
```

```
## [1] 1.684606 3.884894
```

```
# In the original scale of the Y variable the 95% confidence interval is:
c(exp(pi_lower), exp(pi_upper))
```

```
## [1]  5.390326 48.661791
```

We can confirm our calculations in R as follows (results in log scale):

```
conf_interval = predict(lm1.lm, newdata=data.frame(Dose_hormone=40), interval="prediction", level = 0.95
exp(conf_interval)
```

```
##        fit      lwr      upr
## 1 16.19524 5.386925 48.68936
```

Then, using the variable in its original scale we can say that when the hormone dose is 40, the probability that the number of insects caught is between 5.39 and 48.66 is 0.95 (95%).

Question 2

The use of insecticides is beneficial for increasing agricultural production but is a major concern for consumers' advocates and environmentalists. Insecticides protect crops against insect damage, but insecticide use may be harmful to humans.

A horticultural researcher working for New York State extension in Oneida County, in central New York, would like to investigate the relationship between the size of apples and the concentration of a new insecticide (ppm) retained in them. She first applies the insecticide across their experimental orchard following guidelines set forward by USDA. The orchard contains dozens of each apple variety commonly grown in New York. She then randomly selects 41 trees and harvests 1 apple on each of these sampled trees. The amount of insecticide retained by each apple is determined in the lab. The diameter of each apple is also measured. (Data for this problem can be found in the file Hwk2Q2DatSp17.xlsx.)

The average apple in New York State is 6.6 cm across. State regulations for use of this insecticide, which has been shown to be extremely effective and relatively inexpensive, require the average sized apple to contain

less than 2.8 ppm of insecticide, on average. Do these data show that the insecticide can be allowed for use in New York State?

Include the following parts in your answer:

a) Formulation of the research question and choice of the appropriate statistical technique used to answer this question.

The research question is ¿Does the average size apple contains less than 2.8 ppm of insecticide on average?. I think one of the alternatives to deal with these question is to use the simple linear regression. We can model how the size of the apples (X) is related with the concentration of insecticide (Y). Then, we can construct test statistic for the prediction of the mean concentration of insecticide for a average size apple $E(Y|X_i = \bar{X})$. This prediction is a point estimate, so we can construct a t test in the traditional way to test wheter this point estimate is less than 2.8, and then decide wheter or not this insecticide should be allowed in New York State.

b) Notation for the random variable(s) and parameter(s) of interest; define these explicitly. Give the distributional assumptions for your random variable(s) and state all assumptions necessary for the statistical application you intend to use.

We are interested in estimate the equation:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Where: Y_i is the concentration of insecticide in ppm for apple i; X_i is size of apple i in cms; β_0 is the mean value of insecticide's concentration when apple's size is 0; β_1 is the mean increase in insecticide's concentration when the apple's size increase in one cm; and ϵ_i is an error term. In this specification Y_i, β_0, β_1 and ϵ_i are random variables.

The random variables are distributed as follows:

$$Y_i \sim independent N(\beta_0 + \beta_1 X_i, \sigma^2)$$

$$\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$$

$$\hat{\beta_0} \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2})$$

$$\hat{\beta_1} \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2})$$

To estimate the SLR, we assume:

1) Observations are independent ($\epsilon_i \neq \epsilon_j$)
2) The are normally distributed $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$
3) The expected value of Y is a linear function of the variable X: $\mu_i = E(Y_i) = E(Y|X_i) = \beta_0 + \beta_1 X_i$
4) Homocedasticity: ϵ_i have constant variance.
5) Outliers are not driving conclusions

The key random variable that we want to identify is the point estimate

$$E(Y|X_i = \bar{X})$$

This point estimate will tell us what is the expected concentration of insecticide in a average size apple. Then, we are going to construct a test statistic for this point in the traditional way:

$$Test = \frac{pointestimate - estimateunderthenull}{SE(pointestimate)}$$

Where the point estimate will be $E(Y|X_i = \bar{X})$. This test will help us to determine if the mean concentration of insecticide in a average size apple is less than 2.8, and then decide wheter or not this insecticide should be allowed in New York State.

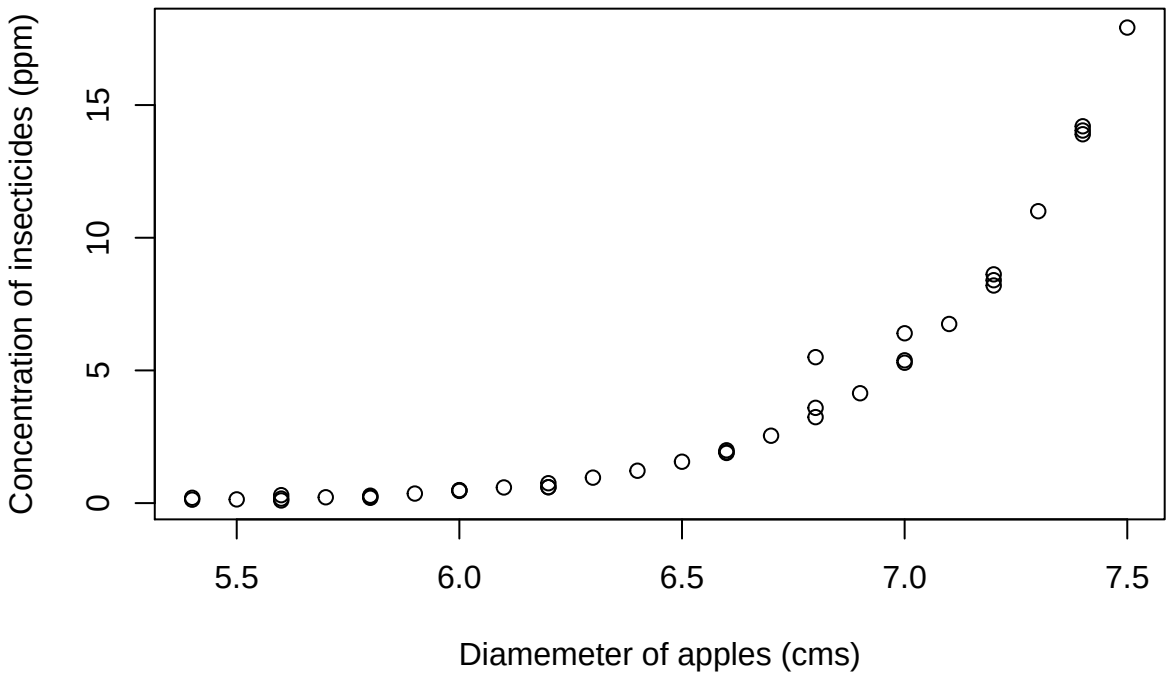
c) Calculations for the analysis. For hypothesis and significance tests, formulate the null and the alternative hypotheses, calculate the value of your test statistic, and then calculate your p-value. For confidence intervals, show and apply the appropriate formula. Use $\alpha = 0.05$ if not otherwise specified.

Initially, we load the data and plot plot the relationship between apple's diameter and concentration of insecticide.

```
library(readxl)
data_apples = read_excel("Hwk2Q2DatSp17.xlsx")
names(data_apples)
```

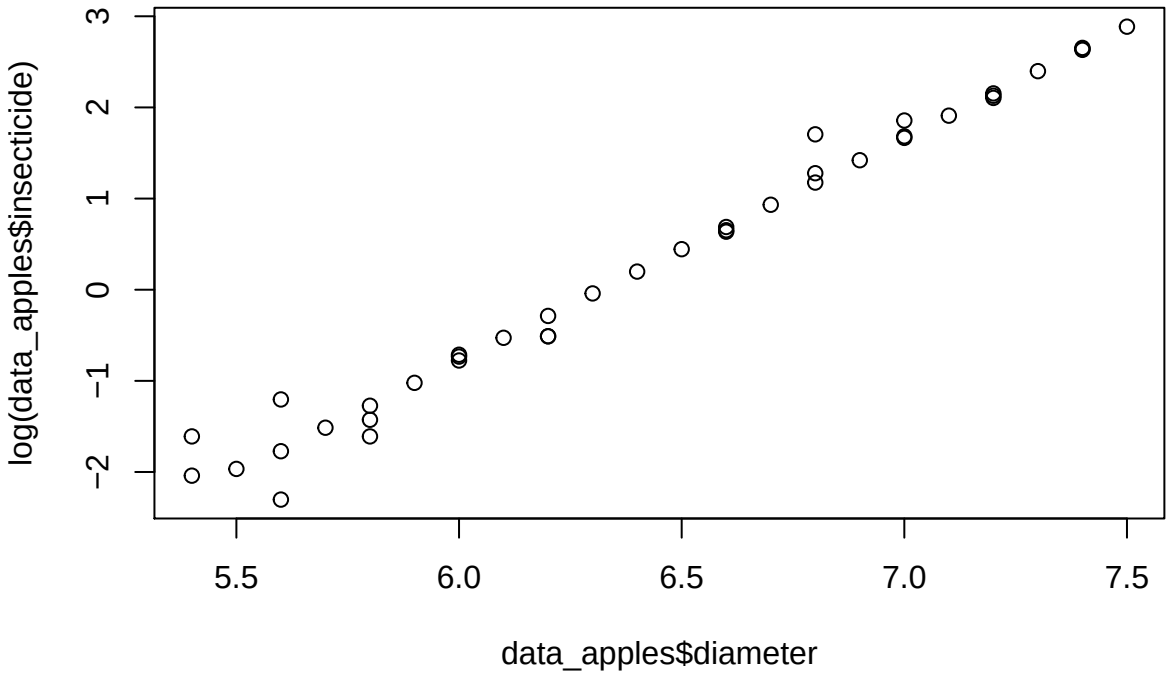
```
## [1] "diameter"    "insecticide"
```

```
plot(data_apples$diameter, data_apples$insecticide, ylab="Concentration of insecticides (ppm)", xlab="Di
```



As we can see the relationship in the variables is no linear, so we transform Y to natural log to linearize the relationship:

```
plot(data_apples$diameter,  log(data_apples$insecticide))
```



Now, we run the SLR to get:

```
lm2.lm = lm(log(insecticide)~diameter, data = data_apples)
summary(lm2.lm)
```

```
##
## Call:
## lm(formula = log(insecticide) ~ diameter, data = data_apples)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57704 -0.03159 -0.01361  0.00227  0.60178
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.32433    0.32368  -47.34   <2e-16 ***
## diameter      2.42836    0.04993   48.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2034 on 39 degrees of freedom
## Multiple R-squared:  0.9838, Adjusted R-squared:  0.9834
## F-statistic:  2365 on 1 and 39 DF,  p-value: < 2.2e-16
```

We are interested in the point estimate $E(Y|X_i = \bar{X})$. In R we can calculate this as follows:

```
B0_1 = -15.32433
B1_1 =  2.42836
mean_x_1= mean(data_apples$diameter)
mean_x_1
```

```
## [1] 6.45122
```

```
E_insect_mean_apple = B0_1 + (B1_1)*(mean_x_1)
E_insect_mean_apple
```

```
## [1] 0.3415534
```

```
# In the original scale of the varible
O_E_insect_mean_apple = exp(E_insect_mean_apple)
O_E_insect_mean_apple
```

```
## [1] 1.407132
```

We can observe that $E(Y|X_i = \bar{X}) = 1.407132$. Then we can define our null and alternative hypothesis as follows:

$$H_0 : E(Y|X_i = \bar{X}) = 2.8$$
$$H_A : E(Y|X_i = \bar{X}) < 2.8$$

Now, to procede with our inference process, we can contruct a test statistic to determine if the average sized apple contain less than 2.8 ppm of insecticide. We can define this test as follows:

$$Statistic = \frac{1.407132 - 2.8}{SE(pointestimate)}$$

Where the $SE(pointestimate)$ is the standard error for the point estimate for the expected value of Y when X is at its mean. We can calculate this value as follows:

$$SE(E(Y|X_i = \bar{X})) = S_e\sqrt{\frac{1}{n} + \frac{\bar{X} - \bar{X}}{(n-1) * (S_x^2)}}$$

In R:

```
SE_res=0.2034
n=41
mean_x= mean(data_apples$diameter)
var_x= var(data_apples$diameter)
SE_point_estimate= SE_res* sqrt((1 + (1/n) + (mean_x-mean_x)/((n-1) * var_x)))
SE_point_estimate
```

```
## [1] 0.2058655
```

Then, our test statistic is:

$$Statistic = \frac{1.407132 - 2.8}{0.2058655} = -6.7660$$

Now we can find the p-value associate with this test statistic as follows:

```
statistic = (1.407132 - 2.8) / (0.2058655)
statistic
```

```
## [1] -6.765913
```

```
p_value = pt(statistic, 39)
p_value
```

```
## [1] 2.249421e-08
```

$p - value = 0.00000002249421 < 0.05$, then we have evidence to say that the mean concentration of insecticide in a average size apple is less than 2.8. We can conclude that the data shows that the insecticide can be allowed for use in New York State.

Then, we can get the confidence interval:

```
predict(lm2.lm, data.frame(diameter=6.6), se.fit = T, interval="confidence")
```

```
## $fit
##         fit       lwr       upr
## 1 0.7028124 0.6368114 0.7688133
##
## $se.fit
## [1] 0.03263029
##
## $df
## [1] 39
##
## $residual.scale
## [1] 0.2034491
```

```
lower=exp(0.6368114)
upper=exp(0.7688133)
c(lower,upper)
```

```
## [1] 1.890443 2.157205
```

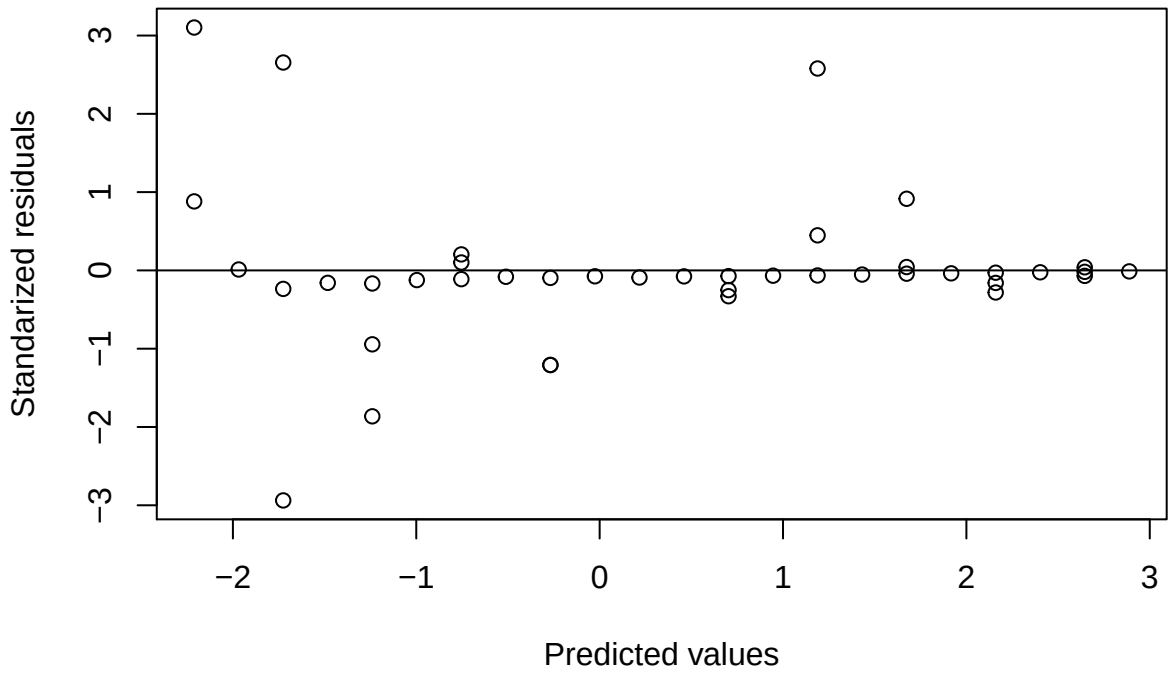
We are 95% confident that the average amount of insecticide remained on average sized apple is within $(exp^{0.64}, exp^{0.77}) = (1.89, 2.16)$.

d) Discuss whether the assumptions stated in Part b) above are met sufficiently for the validity of the statistical inferences; use graphs and other tools where applicable.

1) The assumption of idependence could feasible given the data collection process (SRS).

2) To asses equal variance, we are going to plot the predicted values against the standarized residuals:

```
lm2.stdres=rstandard(lm2.lm)
plot(lm2.lm$fitted.values, lm2.stdres, ylab="Standarized residuals", xlab="Predicted values", abline(0,
```

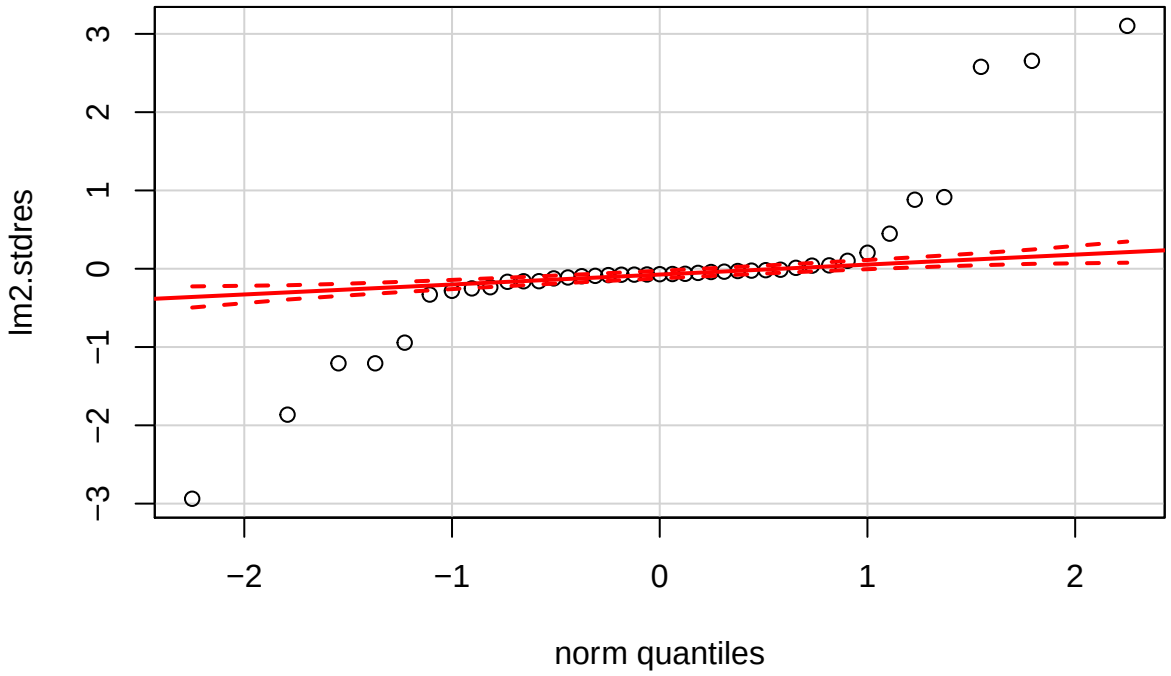


Despite that residuals are centered around zero the variance does not seem to be equal along the predicted values, so the equal variance assumption may be in jeopardy.

3) To assess the normality of the standarized residuals, we can apply two techniques: 1) make a histogram or 2) make a normal quantile comparison plot. We are going to try both:

```
hist(lm2.stdres)
```


```
library(car)
qqPlot(lm2.stdres)
```

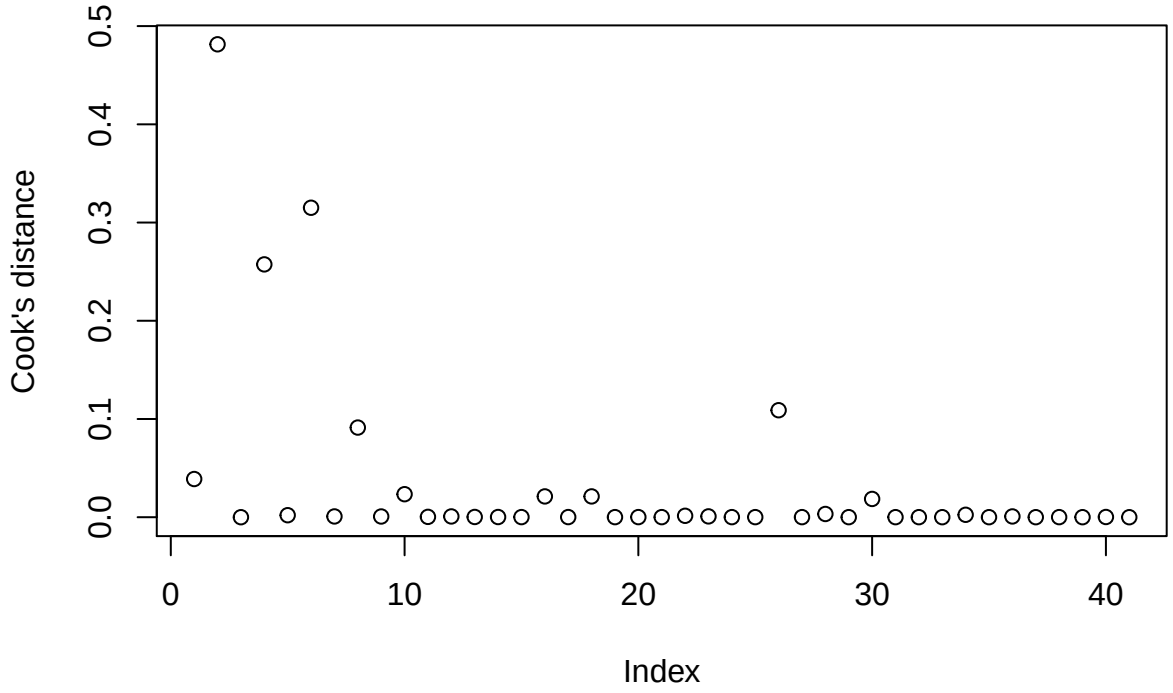


The histogram as well as the qqplot depict that the standarized residuals are not normal, this should affect

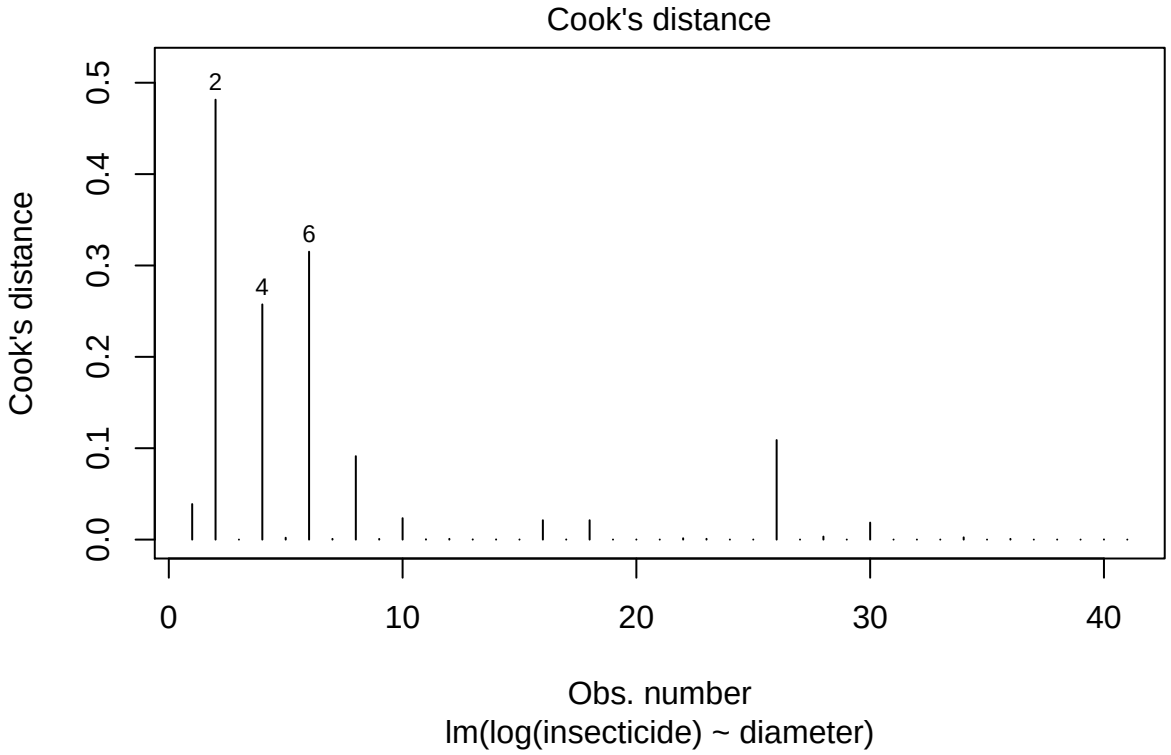
our inference.

4) To assess if we have points highly influential, we can use the cook's distance measure:

```
# cook's distance
cookd = cooks.distance(lm2.lm)
plot(cookd, ylab = "Cook's distance")
```




```
# another way to get the cooks distance
plot(lm2.lm, which = 4)
```



All the values of the cook's distance are less than 1, which means that we do not have a single data point that highly influence the estimated parameters. However, there is a point that despite that is less than one is certainly more influential than the others (observation number 2).

e) Discuss the sampling scheme and whether or not it is sufficient to meet the objective of the study. Be sure to include whether or not subjective inference is necessary and if so, defend whether or not you believe it is valid.

Answer: The sample scheme should have included the variety (type) of the apples. We do not know whether or not some apple varieties react differently to the insecticide. This differectiated effect could be reflected in slopes that change for type of apple, which in turn should affect our estimated slope. I consider that in order to removes the source of variation related with the type of apple, the researcher should have applied a Stratified Random Sampling (strata=type of apple). This sampling method would have ensure a smaller sampling error, which in turn could have improved our inference.

Feedback: However, Even though a typical stratified sampling scheme is performed in collecting the data, the effect is equivalent as a simple random sampling since only one apple is harvested from each selected tree. So all the observations are independent and it is sufficient for the study.

f) State the conclusions of the analysis. These should be practical conclusions from the context of the problem, but should also be backed up with statistical criteria (like a p-value, etc.). Include any considerations such as limitations of the sampling scheme, impact of outliers, etc., that you feel must be considered when you state your conclusions.

The statistical analysis shown that the mean concentration of insecticide in a average size apple is less than 2.8 ppm (p-value < 0.05), then we may say that the insecticide can be allowed to use in New York State. However, some of the assumptions of our statistical model are no met, and this may affect the results of the inference. For instance, the residuals does not seem to be normally distributed, this can certainly affect the efficiency of our estimation (in the sense that our estimators do not have minimun variance and we can improve our efficiency with another method), this situation will be particulary worrying if we had a small simple size (we have a decent sample size of 41). On the other hand, we have righ skewed residuals and highly skewed error distributions tend to generate outliers in the direction of the skew, which in turn compromises

the interpretation of the least square fit. At the same time, we also know that for highly skewed distributions the mean is not a good measure of its center, this situation is particularly important in our case, since we are doing inference for the expected value of y when x is at its mean. Given this considerations, I think we need to repeat the exercise transforming the response variable (Concentration of insecticide) to produce a symmetric error distribution.