BTRY 6020 Homework II Solution

NAME: solution

NETID: solution

DUE DATE: Monday February 13 8:40 am

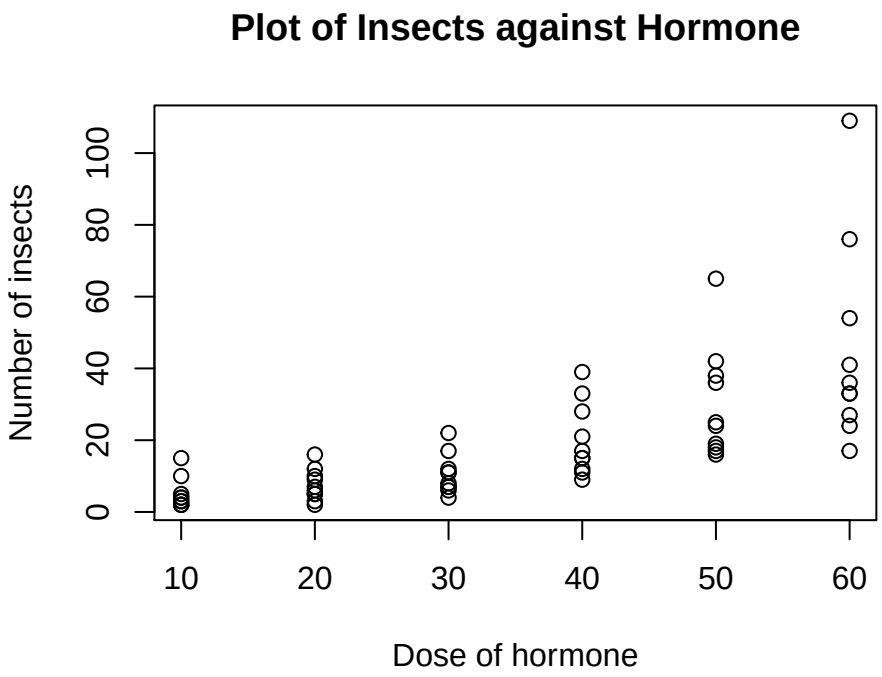
Question 1

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.

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).

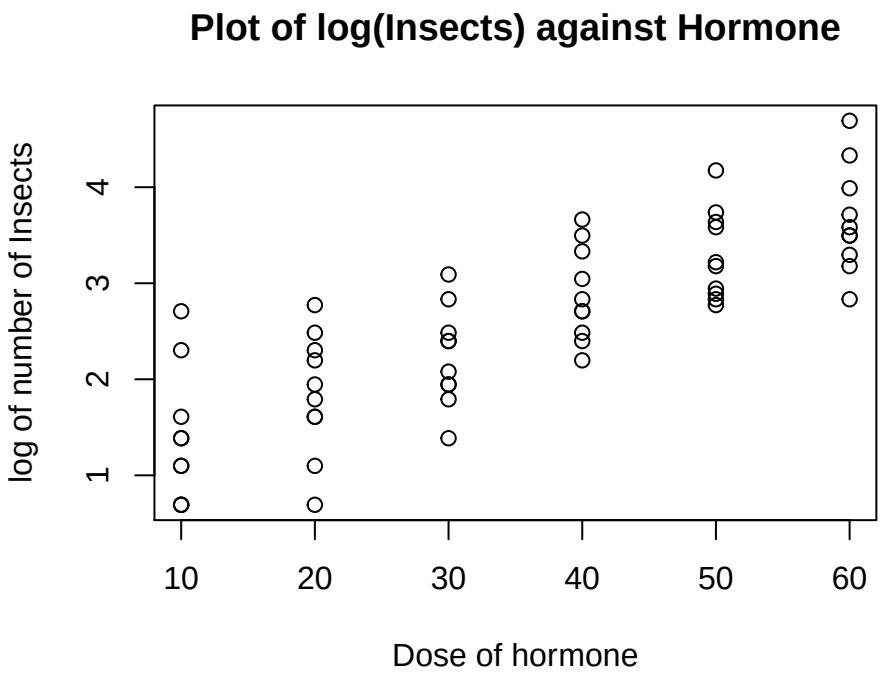
```
library(readxl)
Hwk2Q1DatSp17 <- read_excel("Hwk2Q1DatSp17.xlsx")
# Change the names of the columns
names(Hwk2Q1DatSp17) <- c("ObsNum", "hormone", "insects")
plot(Hwk2Q1DatSp17$hormone, Hwk2Q1DatSp17$insects, xlab="Dose of hormone",
     ylab="Number of insects",
     main = "Plot of Insects against Hormone")
```



The plot shows curvature in the relationship between the two variables, and increasing variance as hormone level increases.

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

```
plot(Hwk2Q1DatSp17$hormone, log(Hwk2Q1DatSp17$insects), xlab="Dose of hormone",
     ylab="log of number of Insects",
     main = "Plot of log(Insects) against Hormone")
```



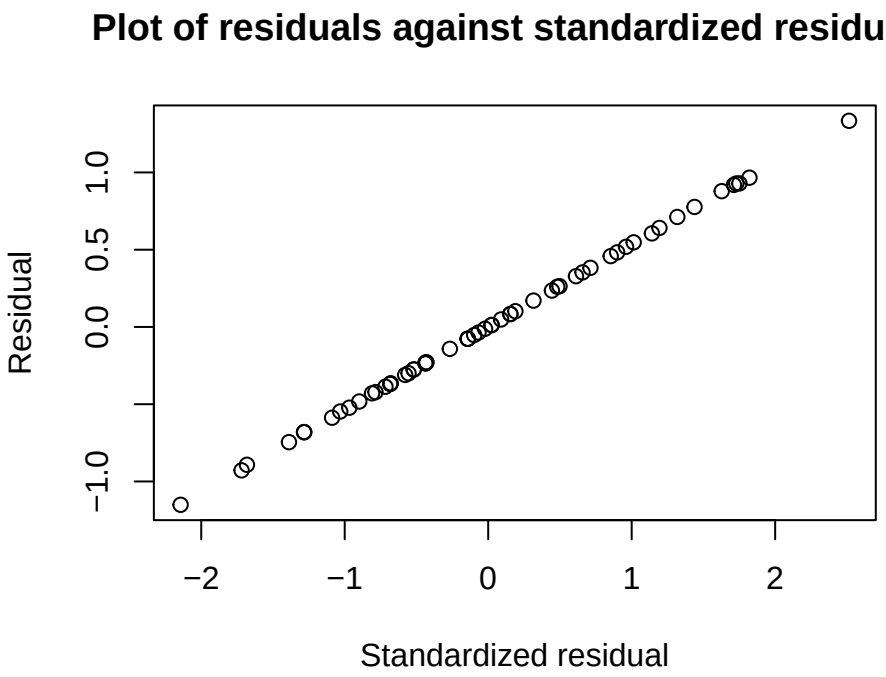
There is a linear relationship between log(Insects) and Hormone and the variance has been stabilized.

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:

```
# Regress log(insects) on hormone
x <- Hwk2Q1DatSp17$hormone
y <- Hwk2Q1DatSp17$insects
insect.lm <- lm(log(y) ~ x)
```

i) Plot the residuals against the standardized residuals. What does this plot reveal?

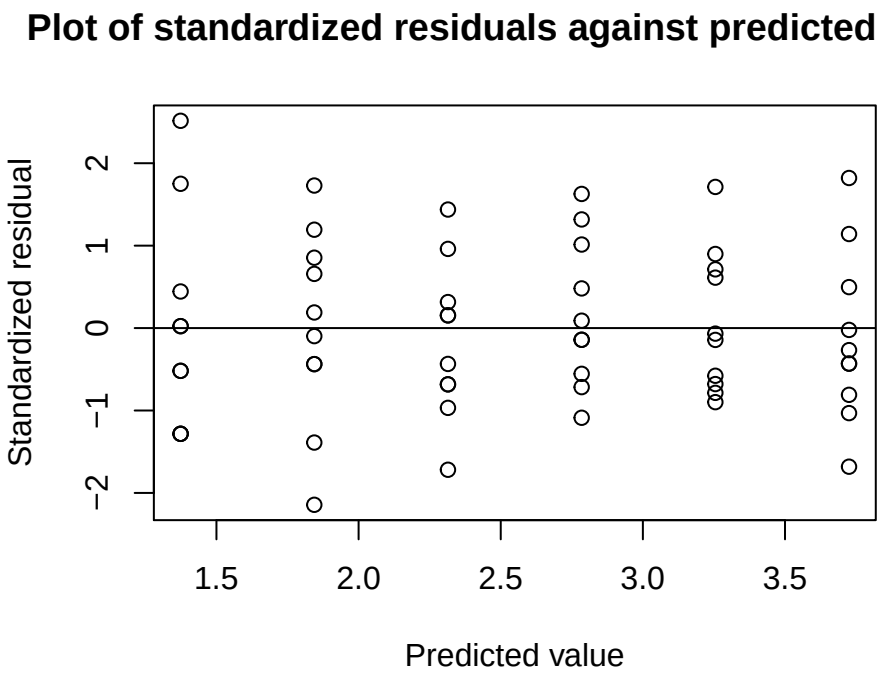
```
insect.stdres <- rstandard(insect.lm)
plot(insect.stdres, insect.lm$resid, xlab="Standardized residual", ylab="Residual",
     main="Plot of residuals against standardized residuals")
```



We see that there is a perfect linear relationship between standardized and typical residuals, implying that residual plots can be made with either and would be exactly the same, except for the y-axis scale. We also notice only two of these standardized residuals lie above +2 or below -2, which is about what we'd 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.

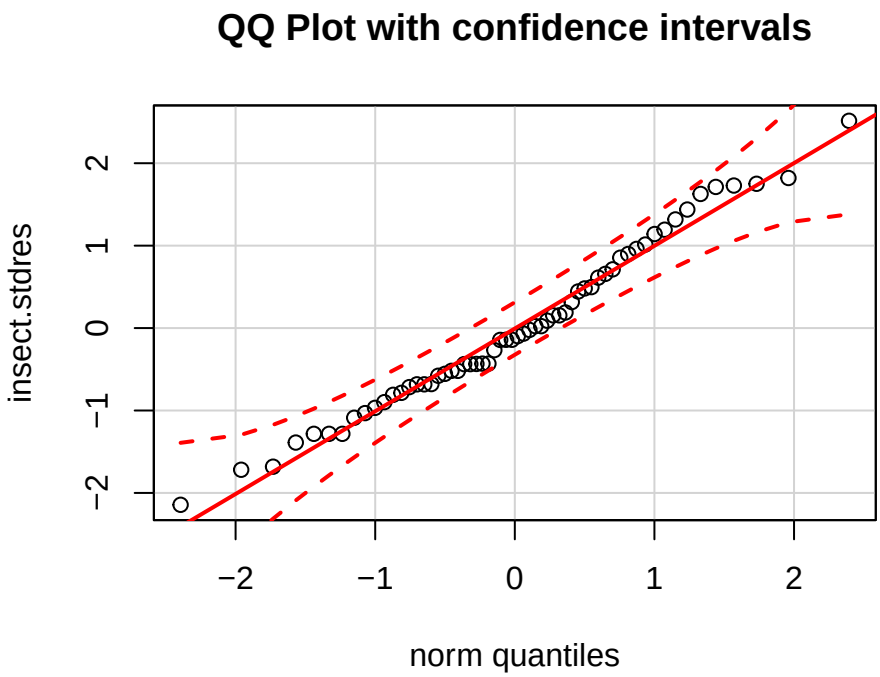
```
plot(insect.lm$fit, insect.stdres, xlab="Predicted value", ylab="Standardized residual",
     main = "Plot of standardized residuals against predicted values")
# Add a horizontal line at y = 0
abline(h=0)
```



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.

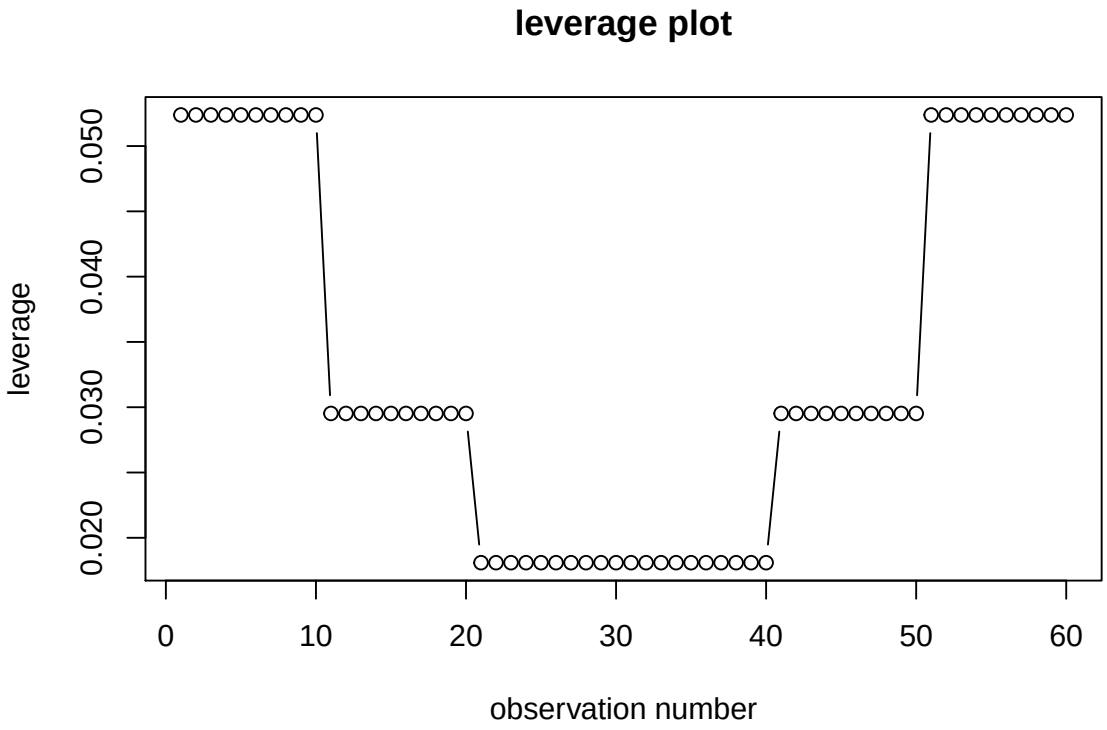
```
library(car)
qqPlot(insect.stdres, main="QQ Plot with confidence intervals")
```



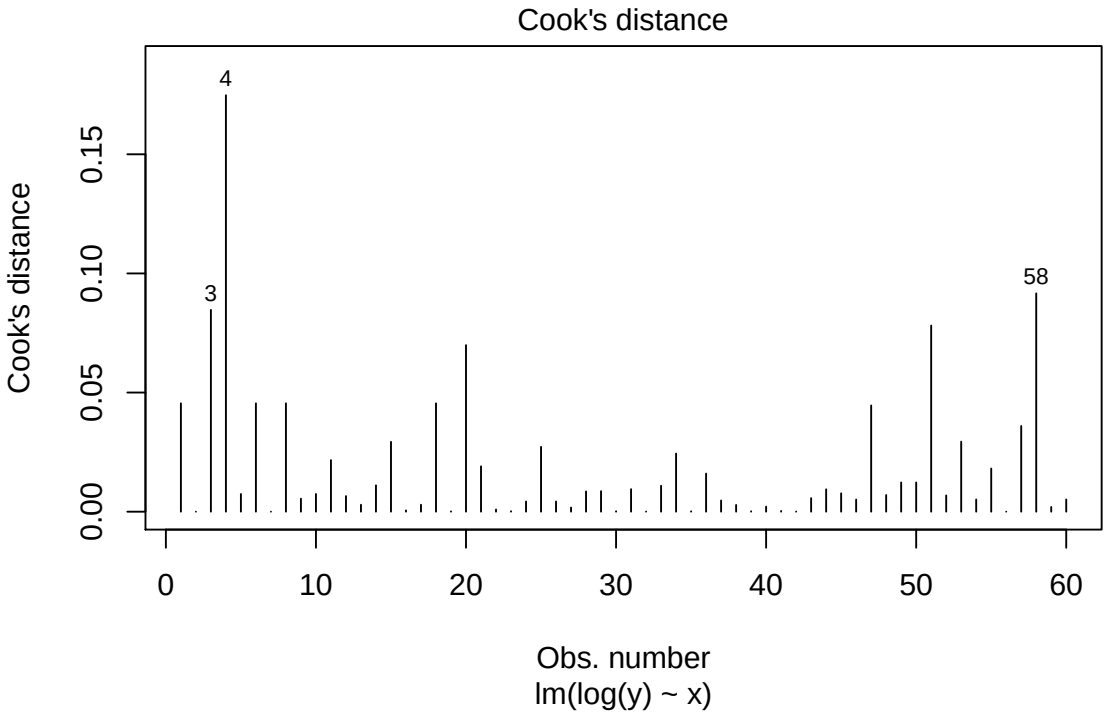
Here the points in the qqplot fall within the confidence bands. So the normality assumption of the (standardized) residuals is satisfied.

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?

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




```
# Cook's distance
plot(insect.lm, which = 4)
```



Since all the leverage are smaller than $\frac{2p}{n} = \frac{4}{60} = 0.067$, there are no high leverage data points. Furthermore, all Cook's distance values are all much less than 1. Therefore, there are no influential data points in this analysis.

d) Compute a 95% confidence interval for the slope of the regression line and carefully state your conclusions.

```
summary(insect.lm)
```

```
##
## Call:
## lm(formula = log(y) ~ x)
##
## 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 ***
## x            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
```

```
# Find 97.5% quantile of t(df=58)
qt(0.975, df=58)
```

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

From the above regression table, a 95% confidence interval for the slope of the regression line is

$$\hat{\beta}_1 \pm t_{58,0.975} \cdot SE(\hat{\beta}_1) = (0.047 \pm 2.0017 * 0.0041) = (0.0388, 0.0553).$$

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.

$H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 > 0$. The test statistic is a t statistic that follows a t_{58} distribution. The test statistic value is $t = 11.416$ from the above summary table. The p-value is less than reported p/2 = 2e-16/2 = e-16. Therefore, we reject H_0 in favor of H_a at $\alpha = 0.05$ and conclude number of insects caught increase with dose.

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.

```
# 95% confidence interval at hormone = 40
predict(insect.lm, data.frame(x=40), interval="confidence")
```

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

```
# 95% prediction interval at hormone = 40
predict(insect.lm, data.frame(x=40), interval="prediction")
```

```
##        fit      lwr      upr
## 1 2.784718 1.683975 3.885461
```

The 95% confidence interval for the mean number of insects caught at hormone = 40 is: $\left(e^{2.6380}, e^{2.9315}\right) = (13.985, 18.756)$. So we are 95% confident that the mean number of insects caught in a trap with a hormone dose of 40 will be between 13.985 and 18.756.

The 95% prediction interval for the number of insects caught at hormone = 40 is: $\left(e^{1.6840}, e^{3.8855}\right) = (5.387, 48.691)$. So the probability that the number of insects caught in a single trap dosed with a hormone level of 40 will be between 5 and 47 is .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?

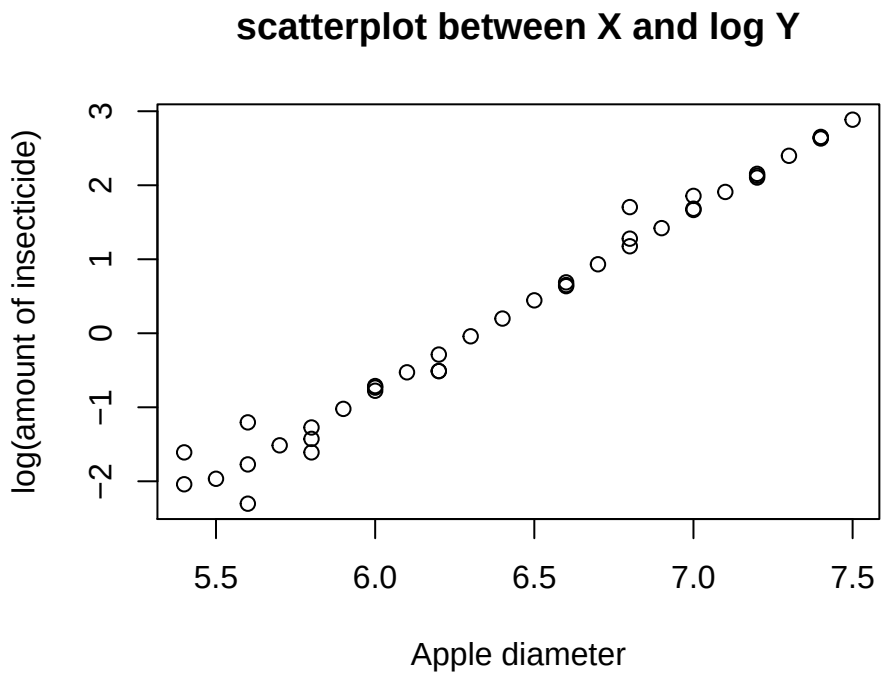
Include the following parts in your answer:

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

The research question is whether an average sized apple (6.6 cm in diameter) contains less than 2.8 ppm of insecticide, on average. In studying the question, we fit a model with dependent variable being the amount of retained insecticide (ppm) and independent variable being the diameter of apple (cm). With the fitted model, we can use hypothesis testing to determine if $\mu_{Y|X=6.6}$ is less than 2.8 ppm.

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.

```
# Import data
Hwk2Q2DatSp17 <- read_excel("Hwk2Q2DatSp17.xlsx")
X <- Hwk2Q2DatSp17$`Diameter of the Apple (X in cm)`
Y <- Hwk2Q2DatSp17$`Insecticide Dose Found (Y, in ppm)`
# scatterplot for checking linearity
plot(X, log(Y), xlab="Apple diameter", ylab="log(amount of insecticide)",
     main="scatterplot between X and log Y")
```



The above scatterplot establishes the linear relationship between log of the amount of retained insecticide, denoted by Y, and the diameter of apple, denoted by X. So a linear model seems appropriate here:

$$log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$$

where ϵ_i is the error term for the ith observation. We make the following assumptions for the model:

i) Observations (and hence residuals) are independent

ii) Normality: $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$

iii) Linearity: Means of $log(Y_i)$ are linearly related to X_i.

iv) Homoscedasticity: ϵ_i have constanct variance.

v) Outliers are not driving our conclusions

vi) 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.

```
# Fit linear model between X and Y
insecticide.lm <- lm(log(Y)~X)
StandResids <- rstandard(insecticide.lm)
predict(insecticide.lm, data.frame(X=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
```

```
qt(0.975, df=39)
```

```
## [1] 2.022691
```

The 95% confidence interval for $\log(\mu_{Y|X=6.6})$ is

$$(\hat{\log y} - t_{df=39,0.975} \times SE(\hat{\log y}), \hat{\log y} + t_{df=39,0.975} \times SE(\hat{\log y}))$$

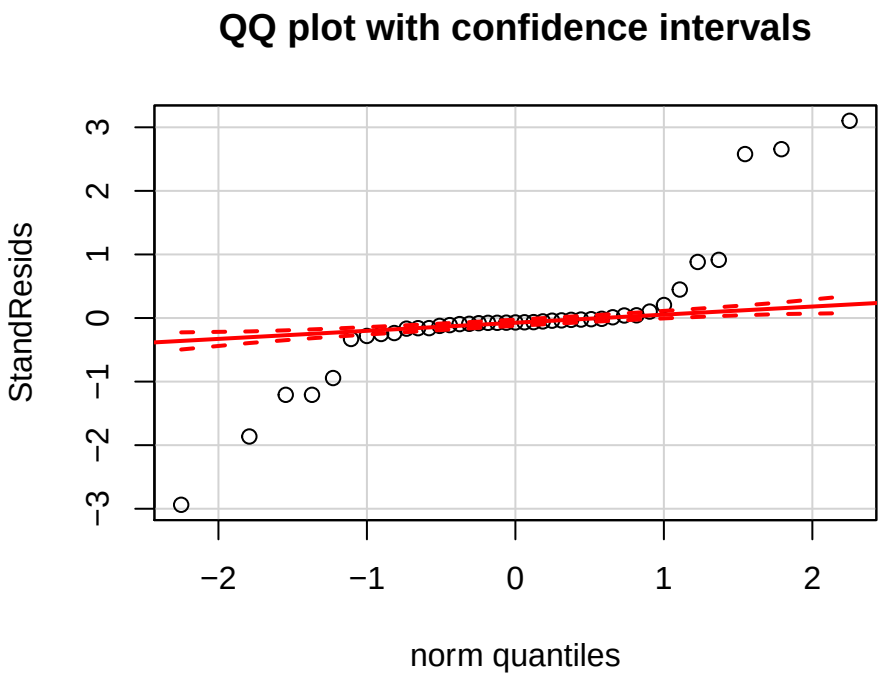
which is (0.64, 0.77) as is provided in the above R output. We are 95% confident that the average amount of insecticide remained on average sized apple is within $(e^{0.64}, e^{0.77}) = (1.9, 2.2)$.

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.

e) Independence is assured from the sampling scheme; insecticide from one apple from a randomly chosen tree should not have any relationship with the amount of insecticide found in another.

ii) Normality of residuals

```
# qqplot for checking normality
qqPlot(StandResids, main="QQ plot with confidence intervals")
```

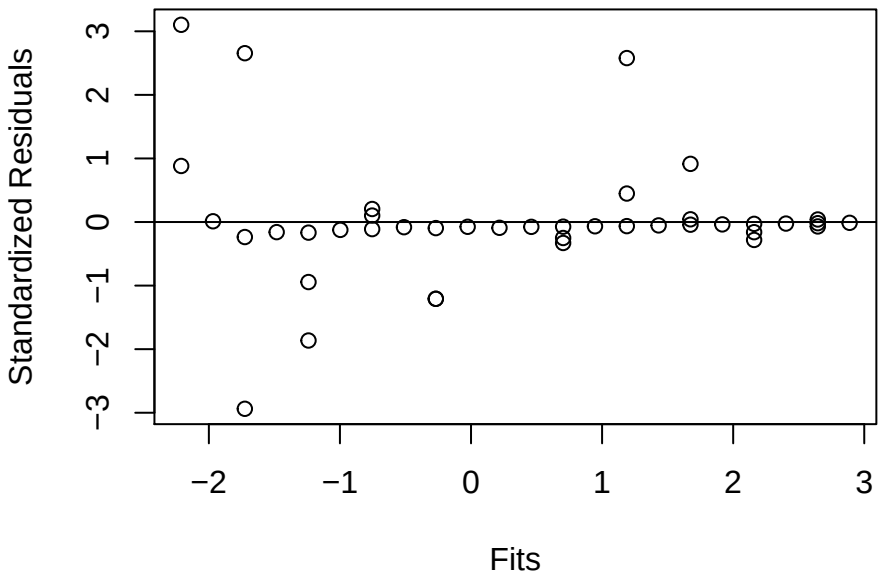


Since many points fall outside of the confidence bands, the normality assumption is not valid from the above qq plot. Points on both ends of the plot are more extreme han they should be, indicating a heavy-tailed distribution of residuals.

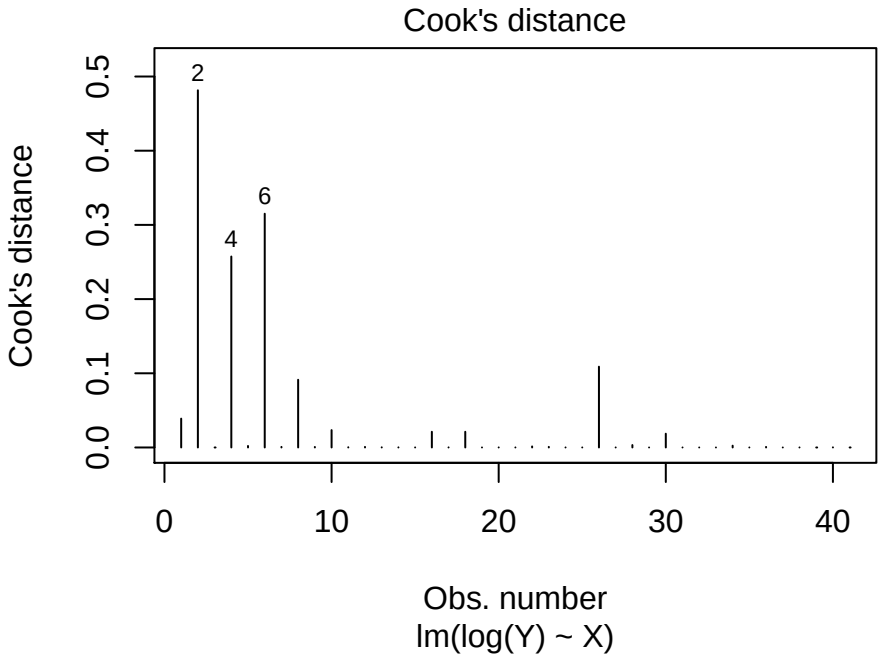
iii) Linearity assumption is verified from the previous scatterplot.

iv) Homoskedasticity: the relatively constant variance around the line seems apparent from the residual plot, with the addition of heavy tails at both ends of this distribution calling this assumption into question.

```
# residuals vs fits for checking homoscedasticity
plot(insecticide.lm$fit, StandResids, xlab ="Fits", ylab="Standardized Residuals")
abline(h=0)
```




```
plot(insecticide.lm, which = 4)
```

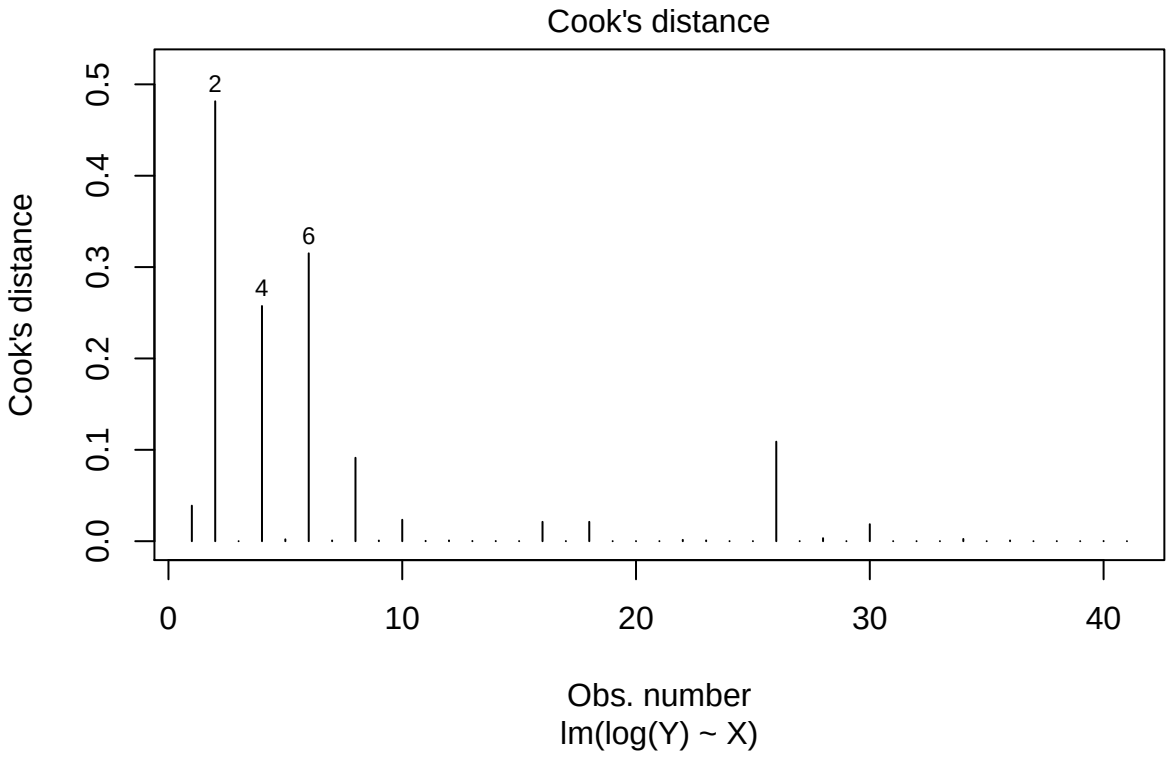


Most of the points in the above residual plot distribute evenly above and below the zero line, so homoscedasticity assumption is valid. However, the plot looks strange as most points are very close to the zero line, with quite a few sticking out above and below it. This heavy-tailed distribution was also evident in the qqPlot of the

standardized resdiuals.

v) Outliers are present in the standardized residual plot (with some exceeding the -3 to +3 threshold). However, let's take a look at a Cook's Distance plot to see if any are influential.

```
plot(insecticide.lm, which = 4)
```



With all Cook's Distances less than 0.5, no particular point seems to be strongly influencing our predicted values

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.

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.

Based on our 95% confidence interval of (1.9, 2.2) ppm insecticide for an average sized apple, we conclude that the average sized apple contains less than 2.8 ppm insecticide. However, as we see above, not all assumptions for our linear model are satisfied; in particular, the observations seem extremely heavy-tailed. The scatterplot shows numerous apples with very high or very low measured insecticide, while the rest fall tightly on a line. This calls into question the measurement of residual insecticide in the apples-it just seems some are way off. With the measurement of residual insecticide being questionable, and assumptions for statistical

inference being invalid, the validity of the above conclusion is questionable. The apparatus for measuring ppm insecticide should be checked and the experiment, and analysis, rerun.