#title: "Assignment 2"

#output: pdf_document

#author: Colm Mooney 20325583

#Q1

bacteria <- read.csv("SharedFiles/ST303/data/Bacteria.csv")

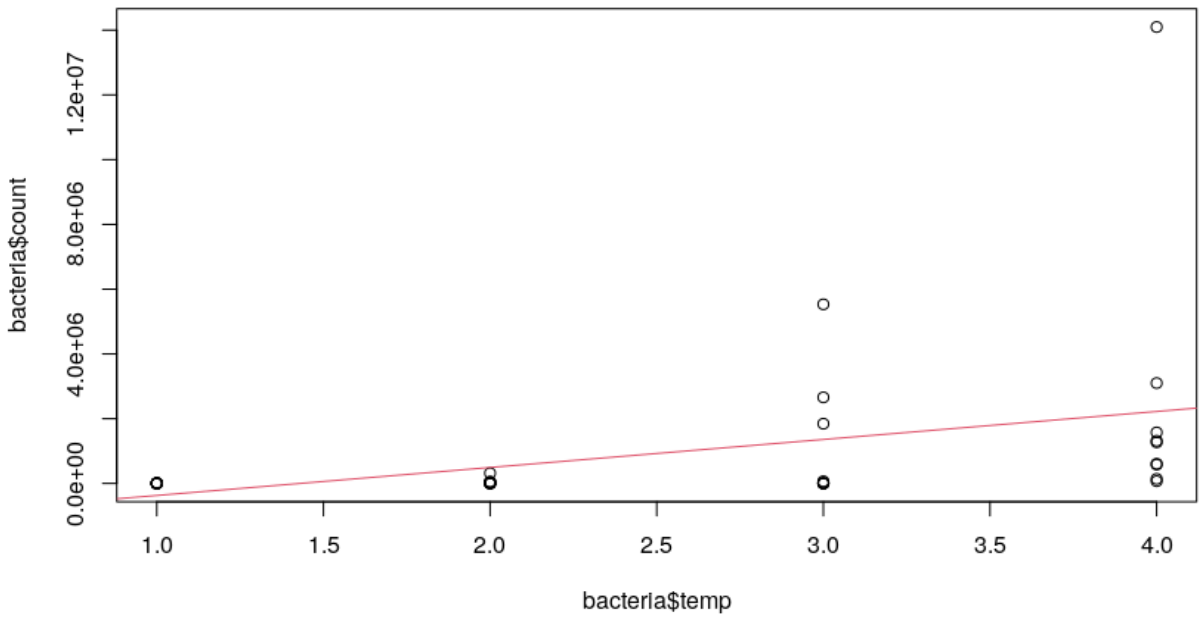
fit <- lm(count ~ temp, data = bacteria)

summary(fit)

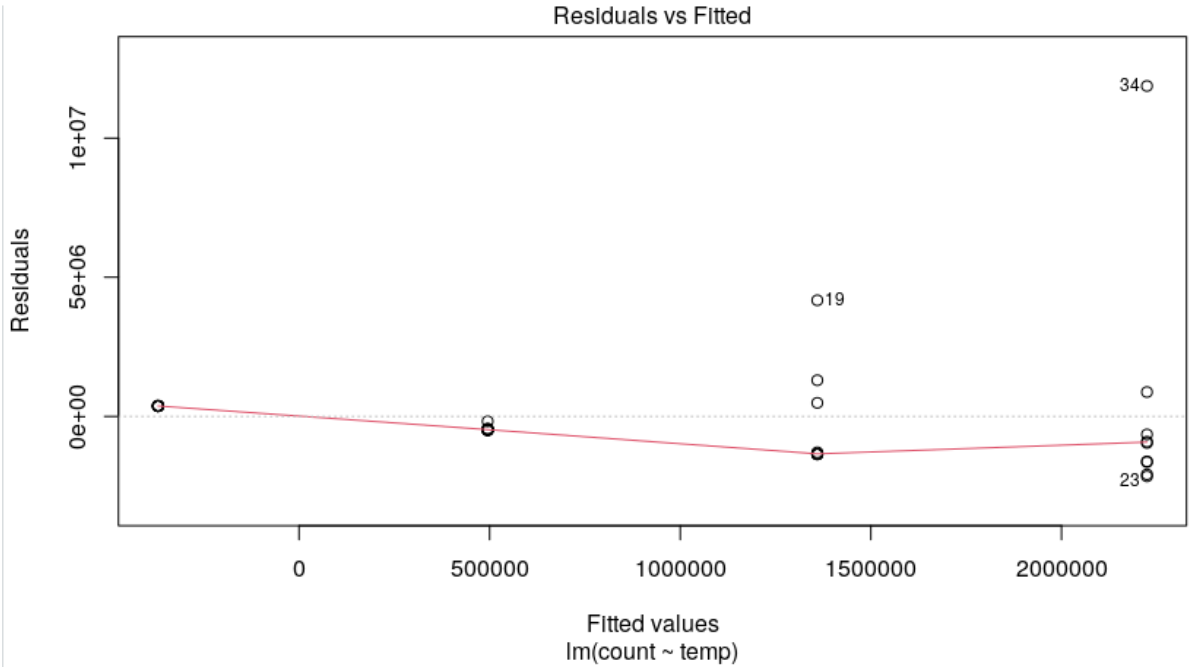
```
Call:
lm(formula = count ~ temp, data = bacteria)

Residuals:
     Min       1Q   Median       3Q      Max 
-2153719 -1305508  -477549   371867 11871844 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1234365     968751  -1.274   0.2112  
temp          864631     353738   2.444   0.0199 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2373000 on 34 degrees of freedom
Multiple R-squared:  0.1495,	Adjusted R-squared:  0.1244 
F-statistic: 5.974 on 1 and 34 DF,  p-value: 0.01985

> 
```

plot(bacteria$temp, bacteria$count)

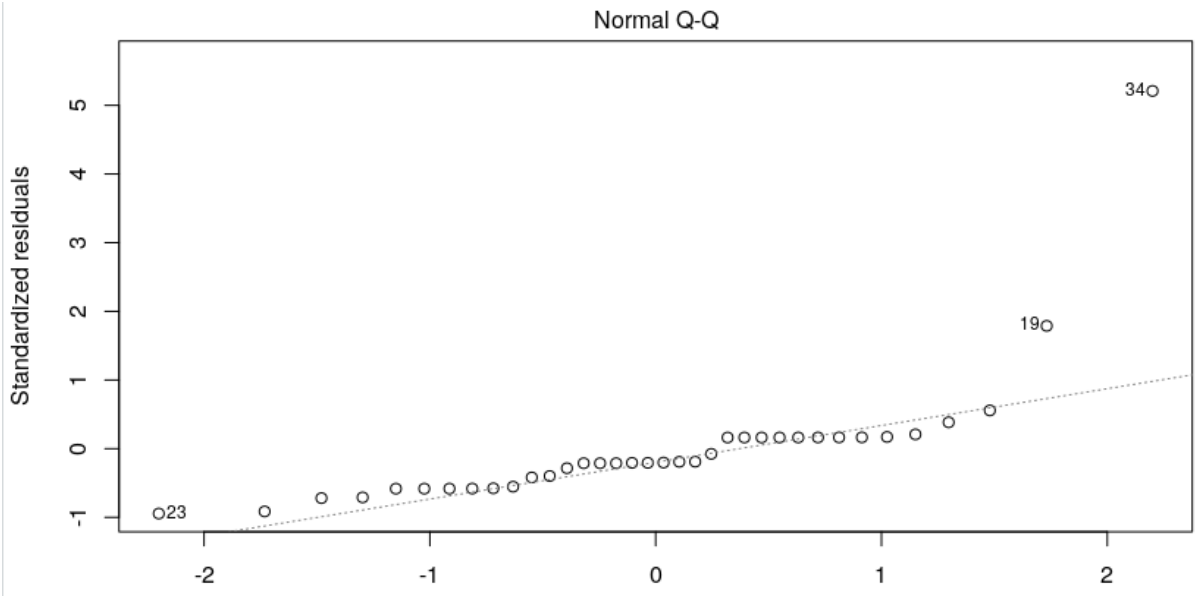
abline(fit, col = 2)



plot(fit, which = 1)



plot(fit, which = 2)



fit2 <- lm(sqrt(count)~temp, data = bacteria)

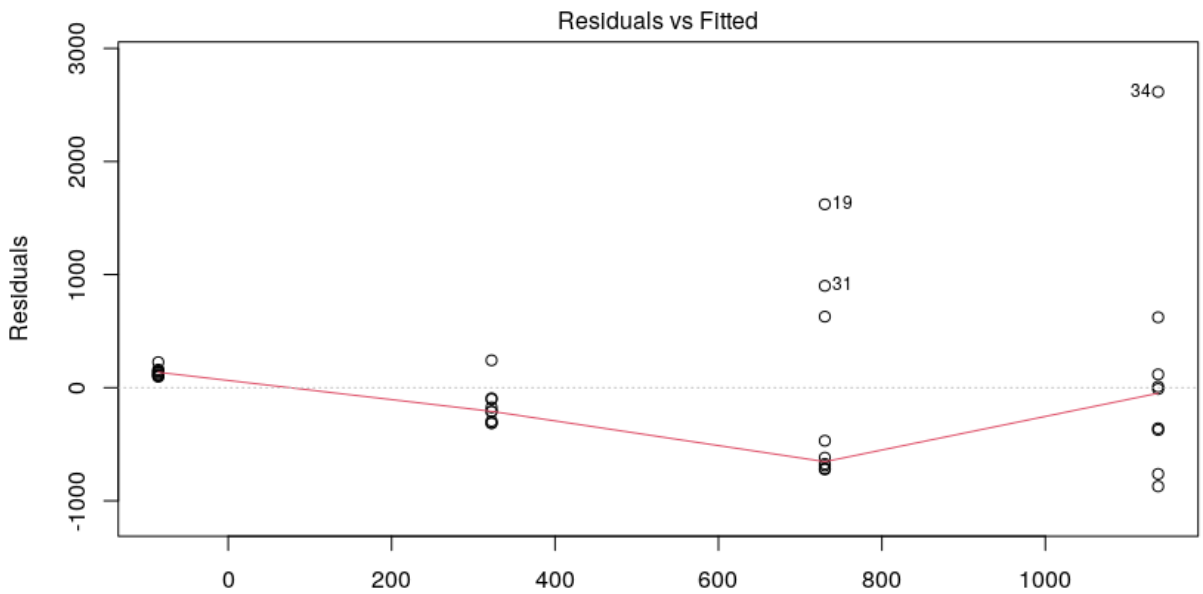
fit3 <- lm(log(count)~temp, data = bacteria)

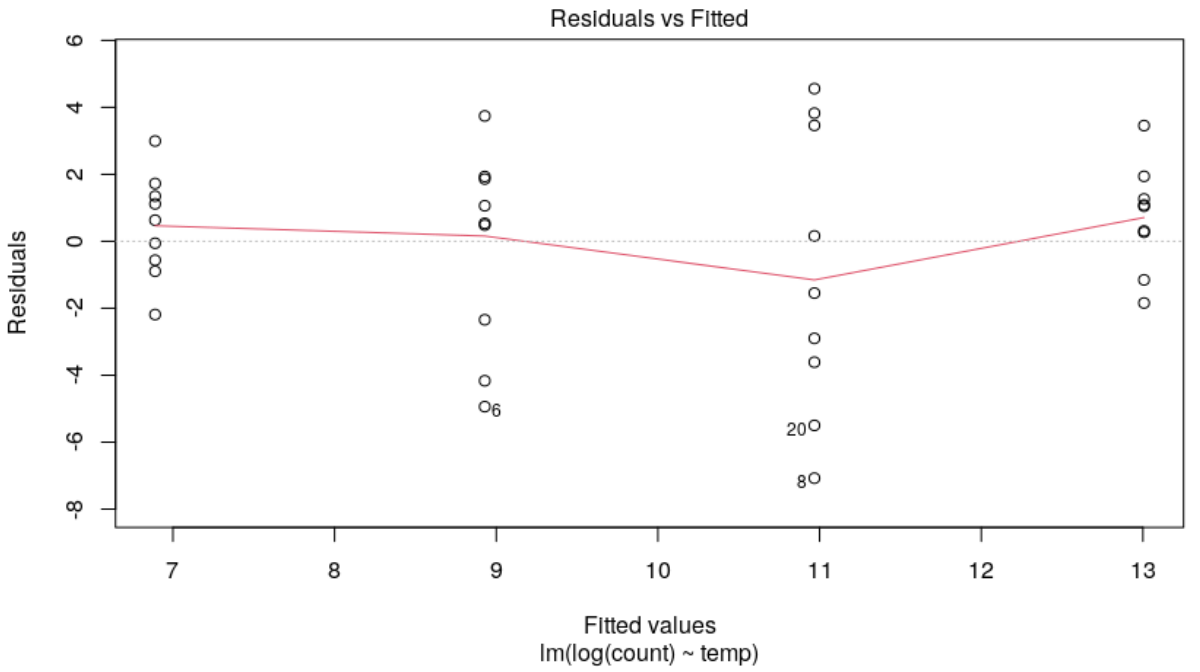
plot(fit2, which = 1)

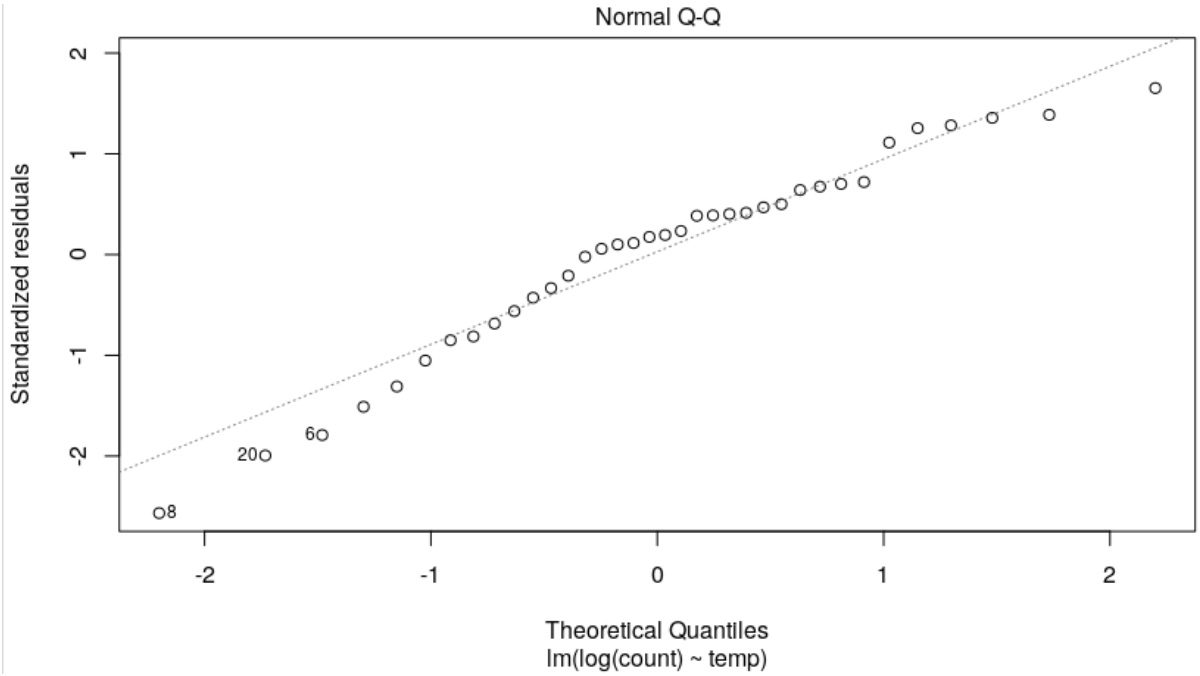
plot(fit2, which = 2) #Same as plot(fit, which = 2) (Not included in graphs.)

plot(fit3, which = 1)

plot(fit3, which = 2)








```
#Commented work.

#mylm2 <- lm(count ~ temp, data = bacteria)

#summary(mylm2)

#anova(mylm2)

#plot( bacteria$temp, bacteria$count)

#abline(lm(bacteria$temp ~ bacteria$count))

#x1 <- bacteria$temp
```

#y1 <- bacteria$count

#fit5 <- lm(y1 ~ x1)

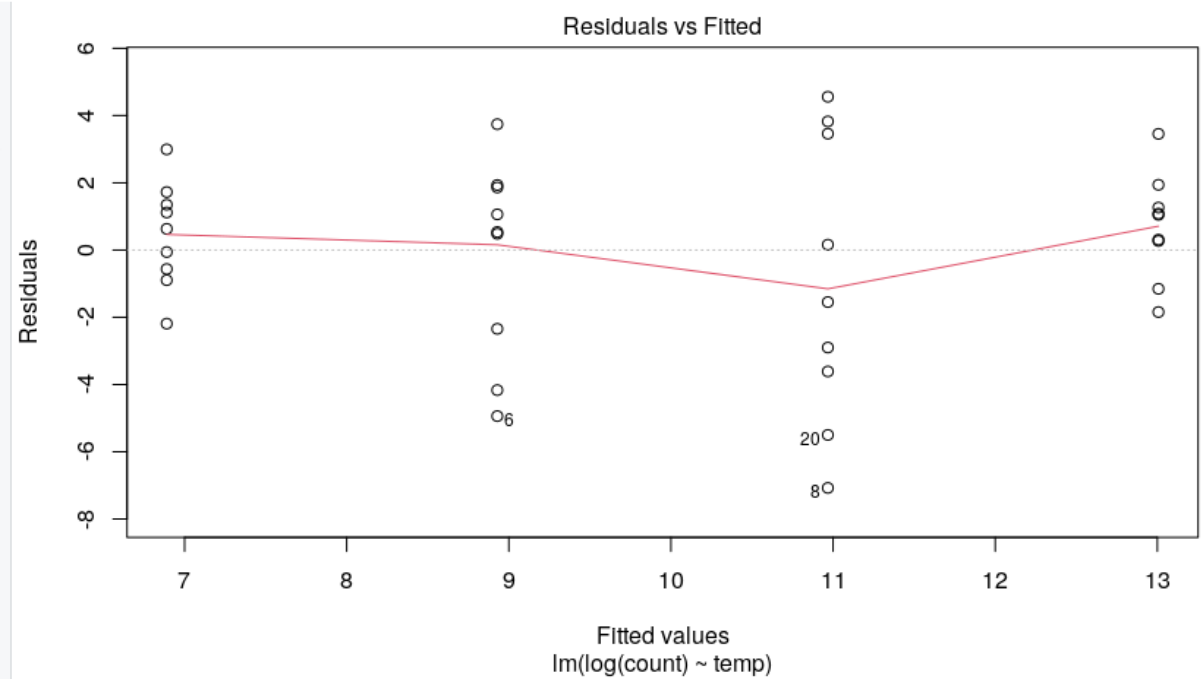
#summary(fit5)

#anova(fit5)

plot(fit3, which =1)

#This isn't a good model. Residuals should always stay around 1 and it isn't doing this in the graph.

#Also, the line is must getting the average location of each Fitted values, & using those 4 mean values instead of doing the line properly.



plot(fit3, which =2)

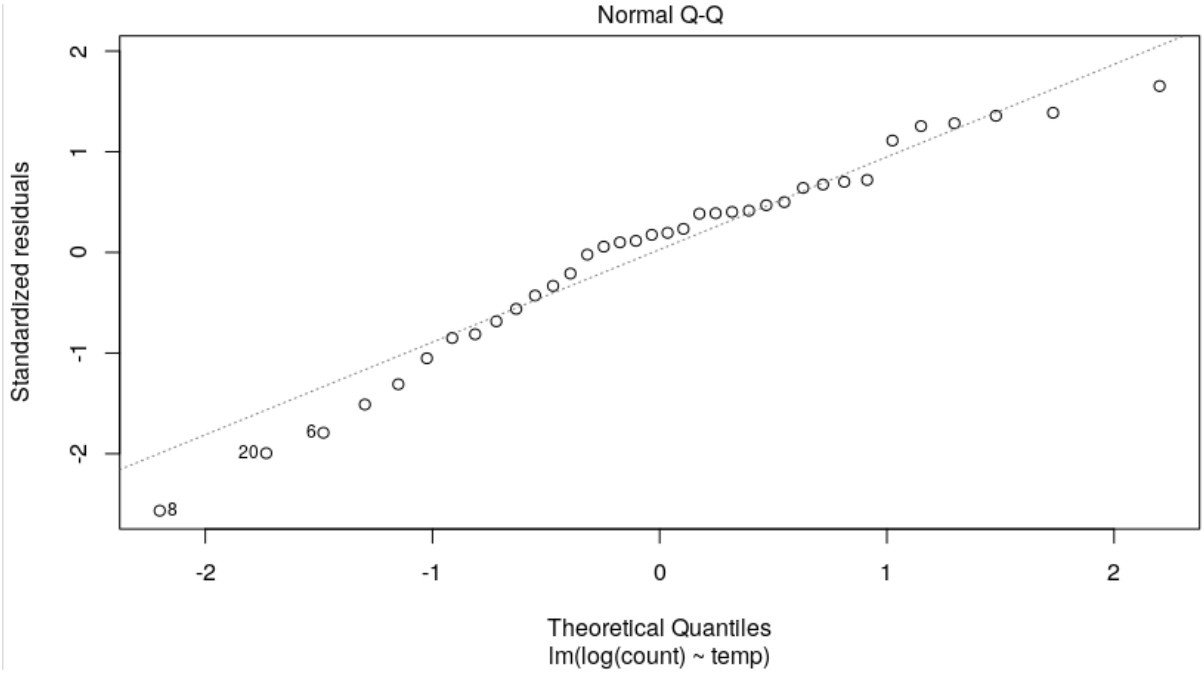
#I believe this is the best model. It is the most linear of the models, the values are all close to 0 and we can see simple linear correlation.

plot(fit, which = 1) #Bad graph, Fitted values not obvious due to large amount, variables are to spread and don't follow line well.

plot(fit, which = 2) #Very good graph, we can easily see some simple linear regression is a clear way.

plot(fit2, which = 1) #Same as plot(fit, which = 1)

plot(fit2, which = 2) #Same as plot(fit, which = 2)

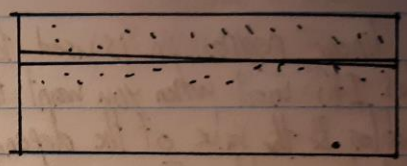


Question 2:

6.1

90/100

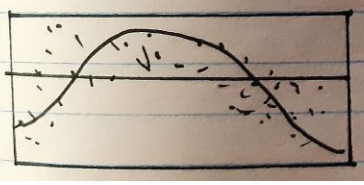
For data set 4, I believe the simple linear regression model is definitely appropriate. ~~I do not see either a positive or a negative relationship between the variables.~~ The residuals stay close to 0 and follow a straight line however, I believe that the single value at (-4,-20) is going to skew the results.



For data set 1, I recommend an exponential transformation, which is a simple algebraic transformation over the simple linear regression (exp)

For data set 3, I recommend the sqrt transformation. This is because high values get compressed & low values become more spread out. Very similar to log transformations but more aggressive in its approach.

For data set 2, I believe the simple linear regression is not appropriate. This is because we cannot predict values outside the range of data with a straight line. This data is more similar to a bell curve like: ∩. This is why I would recommend using a non linear regression model



Question 3:

$Y = 10 + 2x + e$

X	e	2(x)+10	2(x)+10+e
1	-2.2419026	12	9.7580974
2	-0.92071	14	13.07929
3	6.2348333	16	22.2348333
4	0.2820336	18	18.2820336
5	0.5171509	20	20.5171509
6	6.8602599	22	28.8602599
7	1.8436648	24	25.8436648
8	-5.0602449	26	20.9397551
9	-2.7474114	28	25.2525886
10	-1.7826479	30	28.2173521
11	4.8963272	32	36.8963272
12	1.4392553	34	35.4392553
13	1.6030858	36	37.6030858
14	0.4427309	38	38.4427309
15	-2.2233645	40	37.7766355
16	7.1476525	42	49.1476525
17	1.9914019	44	45.9914019
18	-7.8664686	46	38.1335314
19	2.8054236	48	50.8054236
20	-1.8911656	50	48.1088344
21	-4.2712948	52	47.7287052
22	-0.8718997	54	53.1281003
23	-4.1040178	56	51.8959822
24	-2.9155649	58	55.0844351
25	-2.5001571	60	57.4998429
26	-6.7467732	62	55.2532268
27	3.3511482	64	67.3511482
28	0.6134925	66	66.6134925
29	-4.5525477	68	63.4474523
30	5.0152597	70	75.0152597

$9.7580974 = 10 + 2x + e$

$-0.2419026 = 2x + e$

$+\ 2.2419026 = 2x$

$2 = 2x$

$x = 1$

$x = c + dy$

$1 =$

$10 - 2.2419029 - 9.7580974 = 2x$

$\frac{10 - 2.2419029 - 9.7580974}{2} = 1$

(2) $-\frac{10 + e - Y}{2} = X$

$-(10 + e - Y) = 2x$

$-10 - e + Y = 2x$

$-e + Y = 2x + 10$

$Y = 2x + 10 + e$

$Y = 2x + (10 + e)$

$-\frac{10 + e - Y}{2} = \frac{-10 - e + Y}{2} = X$

$\left(\frac{(-10 - e)}{2} + \frac{Y}{2}\right) = X$

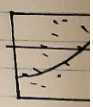
② A simp
resulting
discuss
suggest

simpl
variab
2 var
indepe

For data
With th

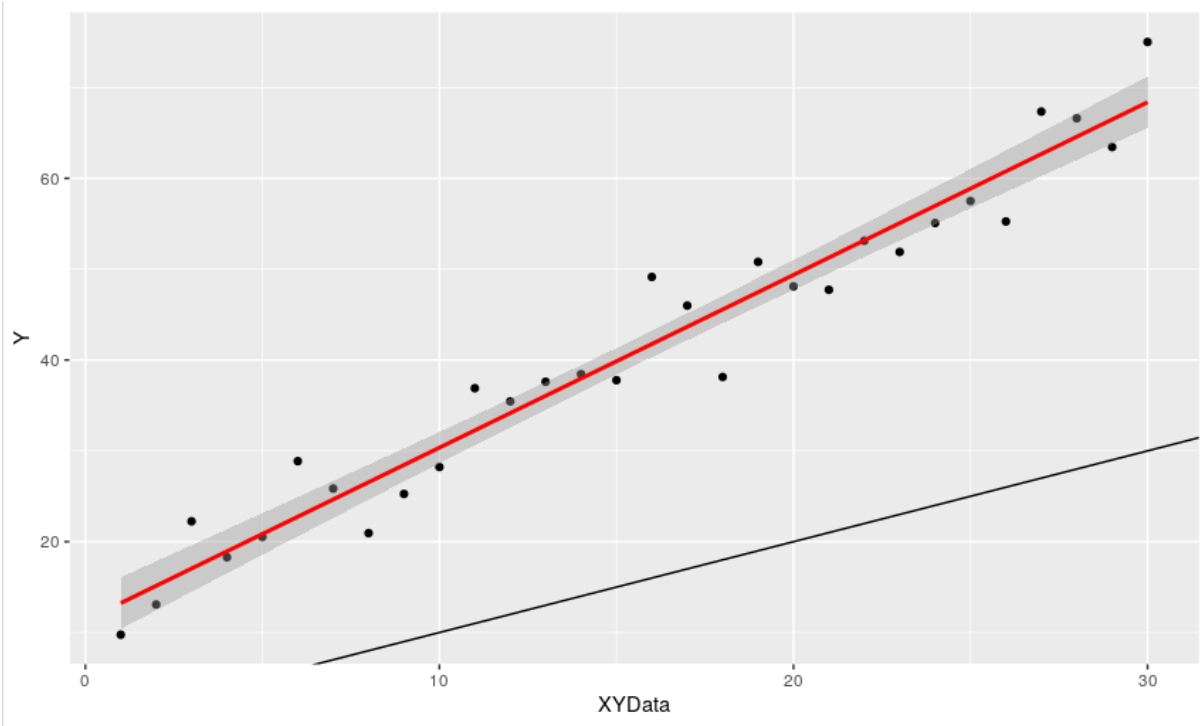
This is
distrib

For dat
because
line. Th
would

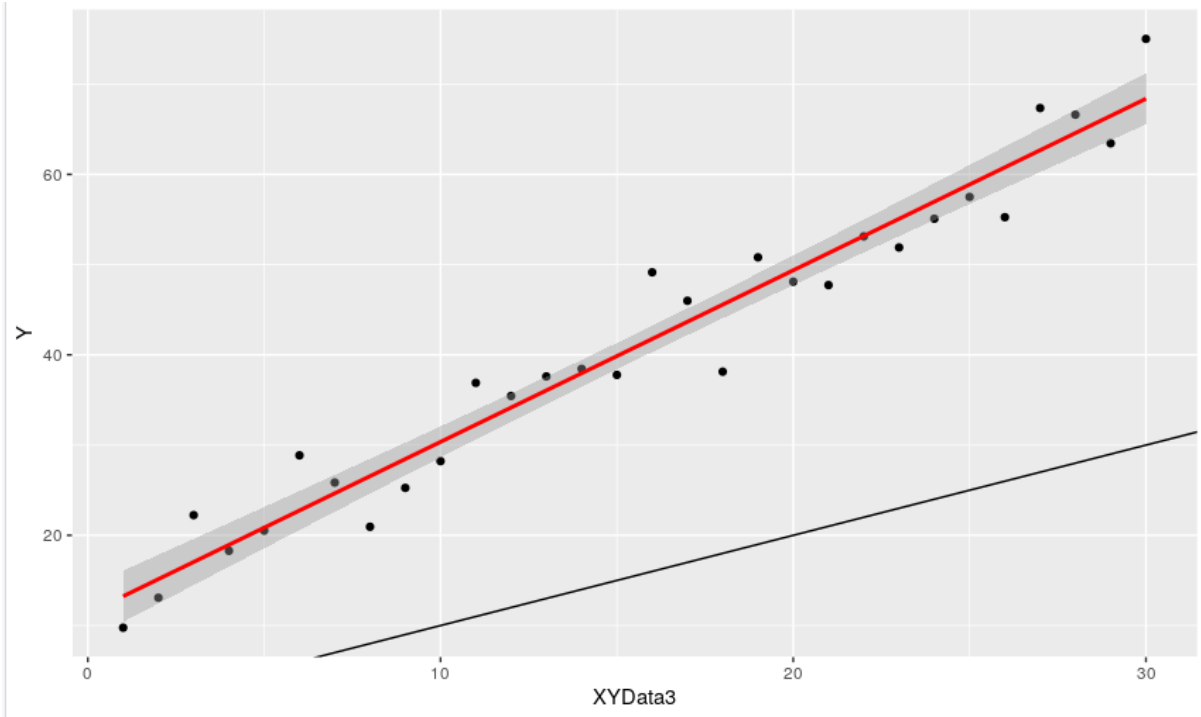


For dat
values

```
set.seed(123)
x <- c(1:30)
X2 <- x * 2
P3 <-X2+10 #bx, e=a
e<- rnorm(30, mean=0, sd=4)
y <- P3 + e #10+2x+e
line2 <- -(10+e-y)/2


XYData <- data.frame(cbind(x,y))
XYData %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_abline() +
  geom_smooth(method = "lm", col = "red") +
  xlab("XYData") +
  ylab("Y")
```




```
#XYData2 <- data.frame(cbind(x,line2))
#XYData2 %>%
 # ggplot(aes(x = y, y = line2)) +
 # geom_point() +
  #geom_abline() +
  #geom_smooth(method = "lm", col = "red") +
  #xlab("XYData2") +
  #ylab("Y")
XYData3 <- data.frame(cbind(y,x))
XYData3 %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_abline() +
  geom_smooth(method = "lm", col = "red") +
  xlab("XYData3") +
  ylab("Y")
```



(c) Yes, the lines are the same.

y is equal to 2x + (10+e)

#It's a matter of finding how much a y is worth equivalent to an x and vice versa, addng in other values to make it consistent.

#Q4

#Given Code

```
library(tidyverse)

library(dplyr)

pollen <- read.table("SharedFiles/ST303/data/pollen.txt", header = TRUE)

head(pollen)

table(pollen)

pollen <- pollen %>% filter(code==1)

pollen_c <- pollen %>%

  filter(duration < 31)


mylm4 <- lm(removed ~ duration, data = pollen) #This is the fit of the model.

summary(mylm4)

anova(mylm4)
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mylm4 <- lm(removed ~ duration, data = pollen) #This is the fit of the model.
> summary(mylm4)

Call:
lm(formula = removed ~ duration, data = pollen)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26437 -0.11327  0.04184  0.10700  0.28321 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.295204   0.053640   5.503 4.17e-06 ***
duration    0.008106   0.002831   2.863  0.00724 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1653 on 33 degrees of freedom
Multiple R-squared:  0.1989,	Adjusted R-squared:  0.1747 
F-statistic: 8.196 on 1 and 33 DF,  p-value: 0.00724

> anova(mylm4)
Analysis of Variance Table

Response: removed
          Df  Sum Sq  Mean Sq F value  Pr(>F)   
duration   1 0.22398 0.223981  8.1959 0.00724 **
Residuals 33 0.90184 0.027328                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
PollenData <- data.frame(cbind(pollen$removed,pollen$duration))

PollenData %>%

 ggplot(aes(x = pollen$removed, y = pollen$duration)) +

 geom_point() +

 geom_abline() +

 geom_smooth(method = "lm", col = "red") +

 xlab("Pollen Data") +

 ylab("Y")
```

#(iii) Yes, the linear regression model seems appropriate. The values are close together, they are close to 0 and follow a linear distribution.

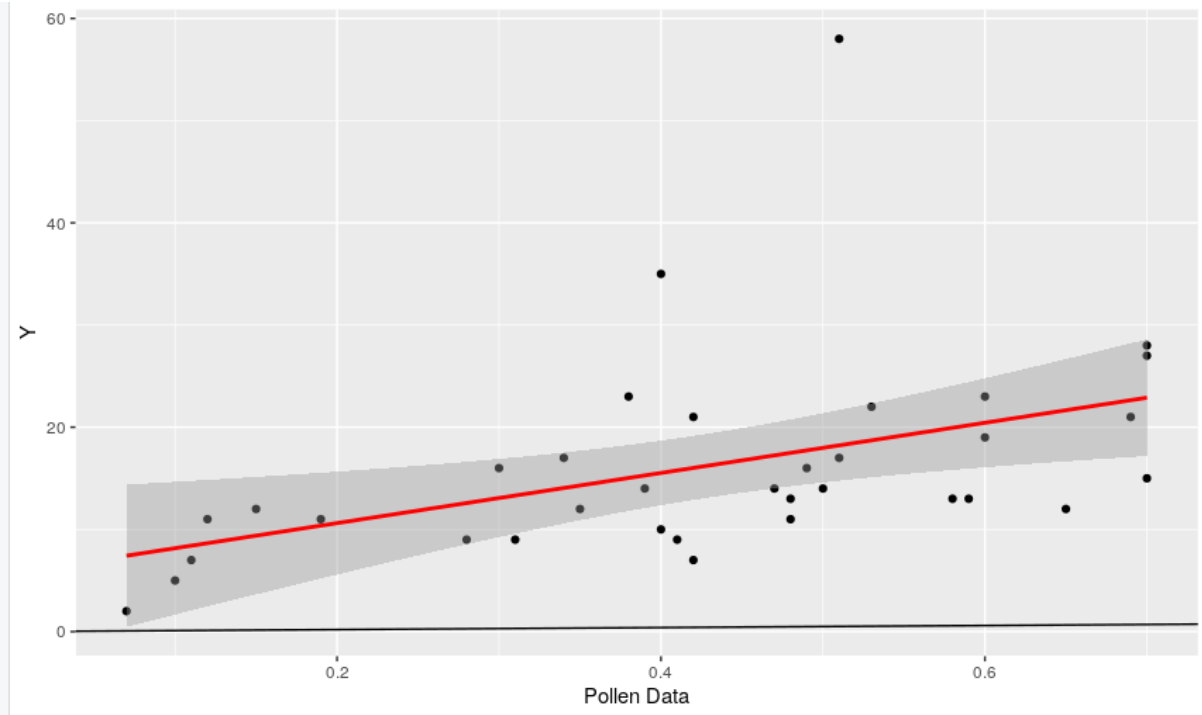
(iv)What problems are evident in the response versus predictor plot?

#That the prediction isn't always acurate. the data shown could actually be following a bell curve but we can't see it.

#There's also the outlier variables that can skew the the line in a direction not suited for the data.

(v)What problems are evident in the residuals versus fitted values plot?

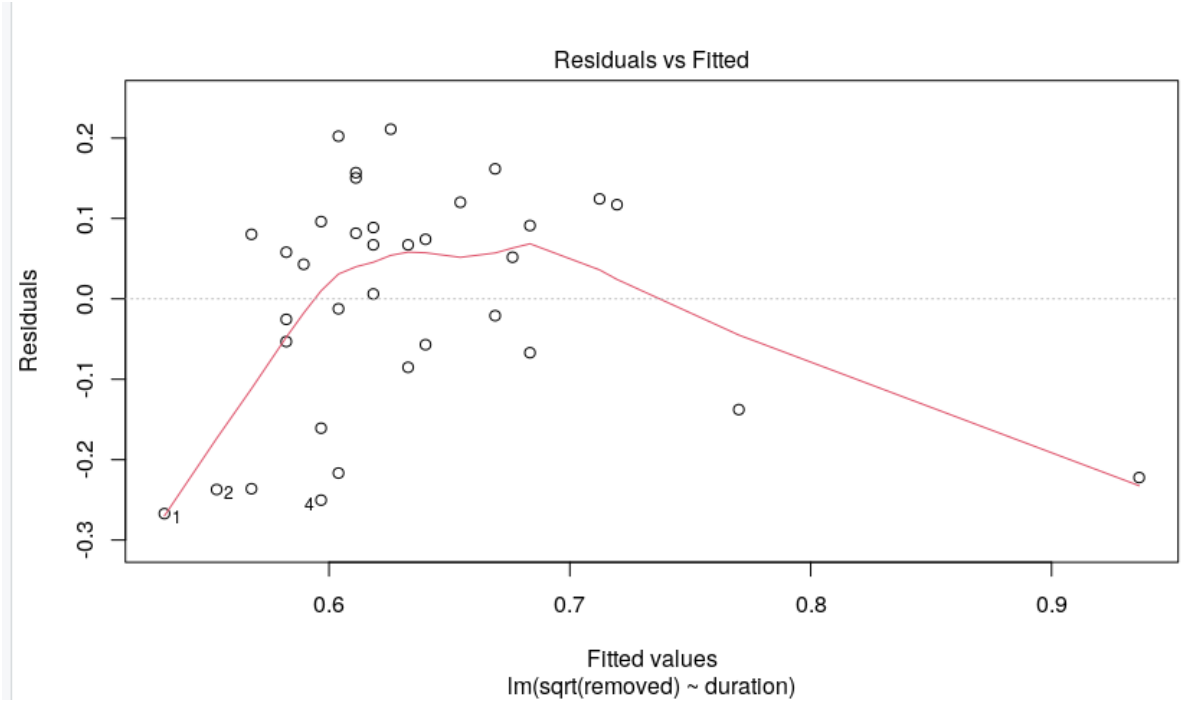
#As I've said before, the outliers can skew the graph.

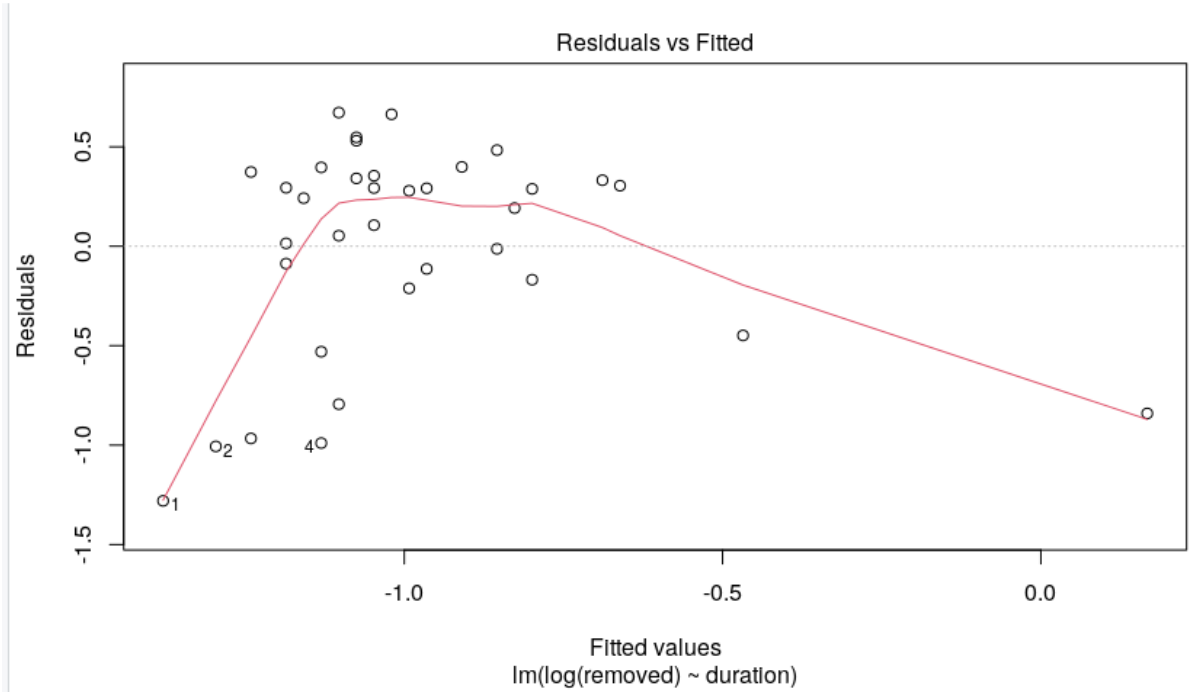


#(b) Do log transformations of Y and / or X help resolve the problems in (a)

```
test2 <- lm(sqrt(removed)~duration, data = pollen)

test3 <- lm(log(removed)~duration, data = pollen)

plot(test2,which = 1)

plot(test3,which =1)
```

#Not particularly, in this case, the data is closer to a bell curve than a striaght line.





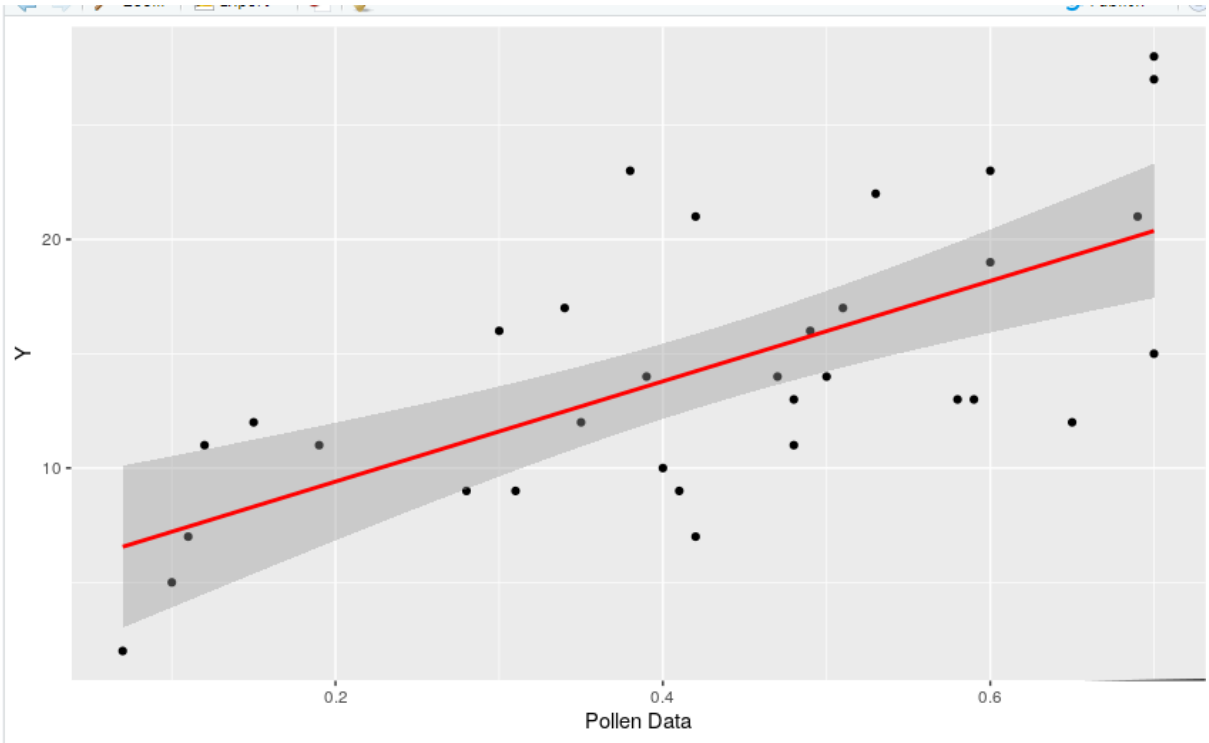
#(c) Try fitting the regression only for those times less than 31 seconds (i.e. excluding the two longest times).

#Does this fit better?

```
Part1=filter(pollen, pollen$duration < 31)

Part1


PollenData2 <- data.frame(cbind(Part1$removed,Part1$duration))
```

```
PollenData2 %>%
 ggplot(aes(x = Part1$removed, y = Part1$duration)) +
 geom_point() +
 geom_abline() +
 geom_smooth(method = "lm", col = "red") +
 xlab("Pollen Data") +
 ylab("Y")
```

#Yes it does fit better. The distribution is closer to the line when the time is less than 31 seconds.



This is the code uninterrupted by graphs/pictures:

#title: "Assignment 1"

#output: pdf_document

#author: Colm Mooney 20325583

#Q1

#Given Code

#(a) Fit a simple linear regression model to these data and provide appropriate graphics to assess the fit of the model.

```
#Identify the issues with the model fit.

#(b) Try appropriate transformations to the response and / or predictor to find a model that is better fit to

#the data than the model in (a). Briefly describe each of the models you fit, discuss how well each model

#fits and indicate which one you deem most appropriate to model the data.


bacteria <- read.csv("SharedFiles/ST303/data/Bacteria.csv")

fit <- lm(count ~ temp, data = bacteria)

summary(fit)

plot(bacteria$temp, bacteria$count)

abline(fit, col = 2)

plot(fit, which = 1)

plot(fit, which = 2)

fit2 <- lm(sqrt(count)~temp, data = bacteria)

fit3 <- lm(log(count)~temp, data = bacteria)

plot(fit2, which = 1)

plot(fit2, which = 2)

plot(fit3, which = 1)

plot(fit3, which = 2)


#Commented work.

#mylm2 <- lm(count ~ temp, data = bacteria)

#summary(mylm2)

#anova(mylm2)

#plot( bacteria$temp, bacteria$count)

#abline(lm(bacteria$temp ~ bacteria$count))

#x1 <- bacteria$temp

#y1 <- bacteria$count

#fit5 <- lm(y1 ~ x1)

#summary(fit5)

#anova(fit5)
```

```
plot(fit3, which =1)

#This isn't a good model. Residuals should always stay around 1 and it isn't doing this in the graph.

#Also, the line is must getting the average location of each Fitted values, & using those 4 mean values instead of doing the line properly.

plot(fit3, which =2)

#I believe this is the best model. It is the most linear of the models, the values are all close to 0 and we can see simple linear correlation.

plot(fit, which = 1) #Bad graph, Fitted values not obvious due to large amount, variables are to spread and don't follow line well.

plot(fit, which = 2) #Very good graph, we can easily see some simple linear regression is a clear way.

plot(fit2, which = 1) #Same as plot(fit, which = 1)

plot(fit2, which = 2) #Same as plot(fit, which = 2)



#Q3

set.seed(123)

x <- c(1:30)

X2 <- x * 2

P3 <-X2+10 #bx, e=a

e<- rnorm(30, mean=0, sd=4)

y <- P3 + e #10+2x+e (y = a + bx),

line2 <- -(10+e-y)/2 #(x = c+ dy)



XYData <- data.frame(cbind(x,y))

XYData %>%

  ggplot(aes(x = x, y = y)) +

  geom_point() +

  geom_abline() +

  geom_smooth(method = "lm", col = "red") +

  xlab("XYData") +
```

```
  ylab("Y")
```

```
#XYData2 <- data.frame(cbind(x,line2))
#XYData2 %>%
 # ggplot(aes(x = y, y = line2)) +
 # geom_point() +
  #geom_abline() +
  #geom_smooth(method = "lm", col = "red") +
  #xlab("XYData2") +
  #ylab("Y")
```

```
XYData3 <- data.frame(cbind(y,x))
XYData3 %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_abline() +
  geom_smooth(method = "lm", col = "red") +
  xlab("XYData3") +
  ylab("Y")
```

#(c) Yes, the lines are the same.

y is equal to 2x + (10+e)

#It's a matter of finding how much a y is worth equivalent to an x and vice versa, addng in other values to make it consistent.

#Q4

#Given Code

```
library(tidyverse)
```

```
library(dplyr)
pollen <- read.table("SharedFiles/ST303/data/pollen.txt", header = TRUE)
head(pollen)
table(pollen)
pollen <- pollen %>% filter(code==1)
head(pollen)
table(pollen)
pollen_c <- pollen %>%
  filter(duration < 31)

#(a)(i) Plot pollen removed versus time spent on flower.
#   (ii) Fit the regression of pollen removed on time spent on flower.
#   (iii)Plot the residuals versus the fitted values. Does the linear regression model seem appropriate?
#   (iv)What problems are evident in the response versus predictor plot?
#   (v)What problems are evident in the residuals versus fitted values plot?


mylm4 <- lm(removed ~ duration, data = pollen) #This is the fit of the model.
summary(mylm4)
anova(mylm4)

PollenData <- data.frame(cbind(pollen$removed,pollen$duration))
PollenData %>%
  ggplot(aes(x = pollen$removed, y = pollen$duration)) +
  geom_point() +
  geom_abline() +
  geom_smooth(method = "lm", col = "red") +
  xlab("Pollen Data") +
  ylab("Y")

#PollenData2 <- data.frame(cbind(pollen$duration,pollen$removed))
```

```
#PollenData2 %>%
#  ggplot(aes(x = pollen$duration, y = pollen$removed)) +
#  geom_point() +
#  geom_abline() +
#  geom_smooth(method = "lm", col = "red") +
#  xlab("Pollen Data") +
#  ylab("Y")
```

#(iii) Yes, the linear regression model seems appropriate. The values are close together, they are close to 0 and follow a linear distribution.

(iv)What problems are evident in the response versus predictor plot?

#That the prediction isn't always acurate. the data shown could actually be following a bell curve but we can't see it.

#There's also the outlier variables that can skew the the line in a direction not suited for the data.

(v)What problems are evident in the residuals versus fitted values plot?

#As I've said before, the outliers can skew the graph.

#(b) Do log transformations of Y and / or X help resolve the problems in (a)

```
test2 <- lm(sqrt(removed)~duration, data = pollen)
test3 <- lm(log(removed)~duration, data = pollen)
plot(test2,which = 1)
plot(test3,which =1)
```

#Not particularly, in this case, the data is closer to a bell curve than a striaght line.

#(c) Try fitting the regression only for those times less than 31 seconds (i.e. excluding the two longest times).

#Does this fit better?

```
Part1=filter(pollen, pollen$duration < 31)
Part1
```

```
PollenData2 <- data.frame(cbind(Part1$removed,Part1$duration))
PollenData2 %>%
  ggplot(aes(x = Part1$removed, y = Part1$duration)) +
  geom_point() +
  geom_abline() +
  geom_smooth(method = "lm", col = "red") +
  xlab("Pollen Data") +
  ylab("Y")
#Yes it does fit better. The distribution is closer to the line when the time is less than 31 seconds.
```

Index of comments

6.1 Is also important to mention why all of them are not appropriate, when they are not.