STAT512 Division 1 HW3

Yi Yang

9/17/2018

1. A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after $1, 3, 5, 7$ and 9 hours. The result is in concentration.csv

a) Fit a linear regression function.
b) Complete the ANOVA table, what is the SSE?
c) Perform the F test to determine whether or not there is lack of fit of a linear regression function; use a significant level of 0.025. Compute the test statistic, reject region, and estimate the p-value. Is the model lack of fit for the data?
d) Perform the diagnostics on the data
   i. Plot the dependent variable versus the explanatory variable and comment on the shape and any unusual points.
   ii. Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.
   iii. Examine the distribution of the random error. What do you conclude?
   iv. Do we need to do a transformation on X or Y? Why or why not?
e) Using the automated Box-Cox procedure, determine which transformation of Y would be appropriate (if any)?
f) Using the appropriate transformation, fit a new model and repeat the diagnostics in part d).
g) With back-transformation, compute and interpret the confidence interval for $\hat{Y}_h$ when $X_h = 7.5$.

Solution:

(a) Fit the line as follows

```
concentration.df <- read.csv('data/concentration.csv',header = TRUE)
names(concentration.df)
```

```
## [1] "Y" "X"
```

```
X <- concentration.df$X
Y <- concentration.df$Y
concentration.mod <- lm(Y~X, concentration.df)
summary(concentration.mod)
```

```
##
## Call:
## lm(formula = Y ~ X, data = concentration.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5333 -0.4043 -0.1373  0.4157  0.8487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
## X            -0.3240     0.0433  -7.483 4.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4743 on 13 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.7971
## F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06
```

The fitted linear regression function is, and the fitted line is plotted.

$$\hat{Y} = -0.3240X + 2.5753$$

(b) The ANOVA table is listed as follows, R is used to compute these values. SSE is error sum squared, which is $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 2.9247$.

Source	degree of freedom	Sum of Squares	Mean Square	F-value
Model	1	12.5971	12.597	55.994
Error	13	2.9247	0.225	
Lack of fit	3	2.7673	0.9224	58.603
Pure error	10	0.1574	0.01574	
Corrected Total	14	15.5218		

```
# Reduced model
anova(concentration.mod)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X          1 12.5971  12.597  55.994 4.611e-06 ***
## Residuals 13  2.9247   0.225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Full model
concentration.mod.full <- lm(Y~as.factor(X),concentration.df)
anova(concentration.mod,concentration.mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ as.factor(X)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     13 2.9247
## 2     10 0.1574  3    2.7673 58.603 1.194e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(1-0.025,3,10)
```

```
## [1] 4.825621
```

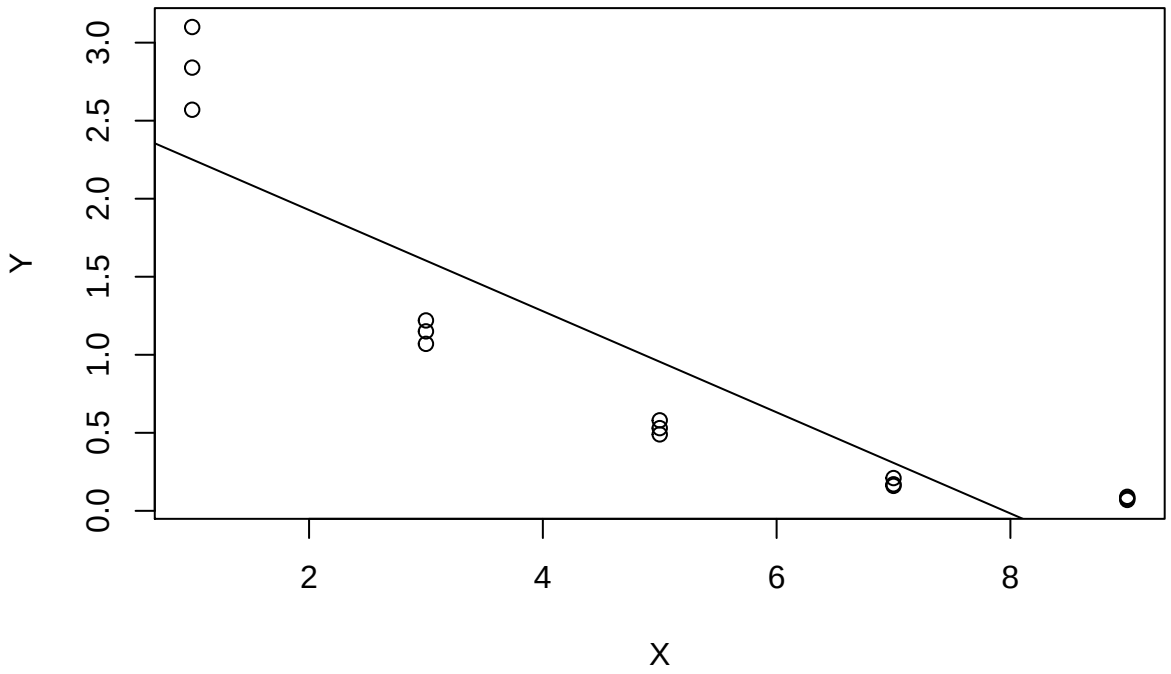
(c) The test hypothesis is drawn as

$$H_0 : E(Y) = \mu = \beta_0 + \beta_1 X, \quad H_a : E(Y) = \mu \neq \beta_0 + \beta_1 X$$

The test statistic is given by $F = 58.603$, rejection region is $F > 4.8256$. The p-value is given by $Pr = 1.194 \times 10^{-6} < \alpha = 0.025$, which means the null hypothesis can be rejected. That is, the current linear model does not fit the data.

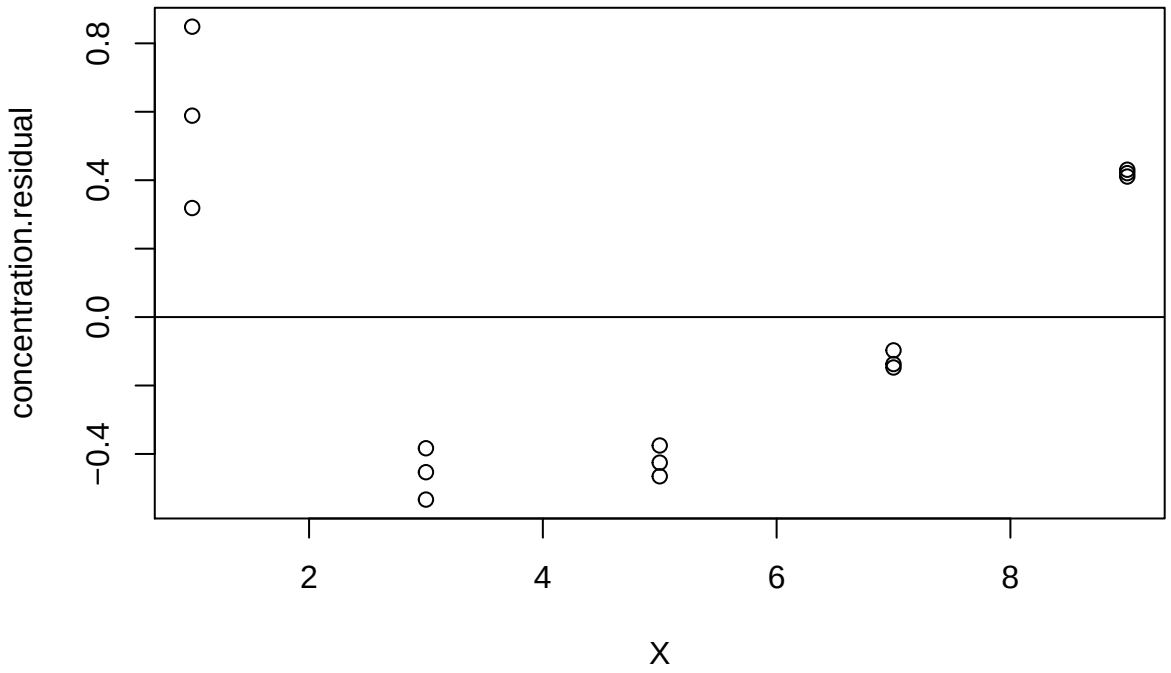
(d.i.) The scatter plot is given below. By observing the distribution of the scatter points, it can be infered that there is a lack fit of linear regression, the error does not have a constant variance and it may not be normally distributed.

```
plot(X,Y)
abline(concentration.mod)
```



(d.ii.) Residual plot is given below. From the residual plot, we find that the linear regression model is not appropriate.The error does not have a constant variance. There seems to exist several outliers at level $X = 1$.

```
concentration.residual <- residuals(concentration.mod)
plot(X,concentration.residual)
abline(h=0)
```

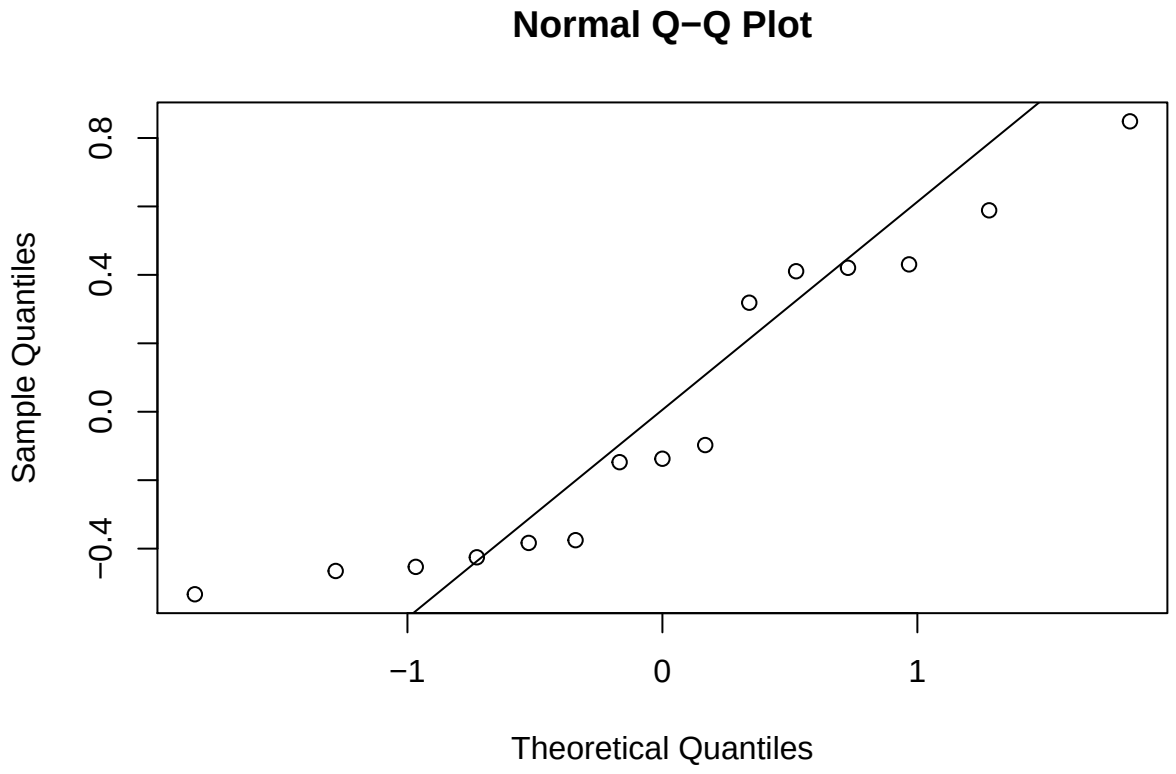


(d.iii) Let us conduct Shapiro test and plot the probability plot to test the normality of error terms. From the test result and the probability plot, the distribution of random errors can be treated as normal distribution since pvalue is greater than significance level $\alpha = 0.05$.

```
shapiro.test(concentration.residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  concentration.residual
## W = 0.88679, p-value = 0.05998
```

```
qqnorm(concentration.residual)
qqline(concentration.residual)
```



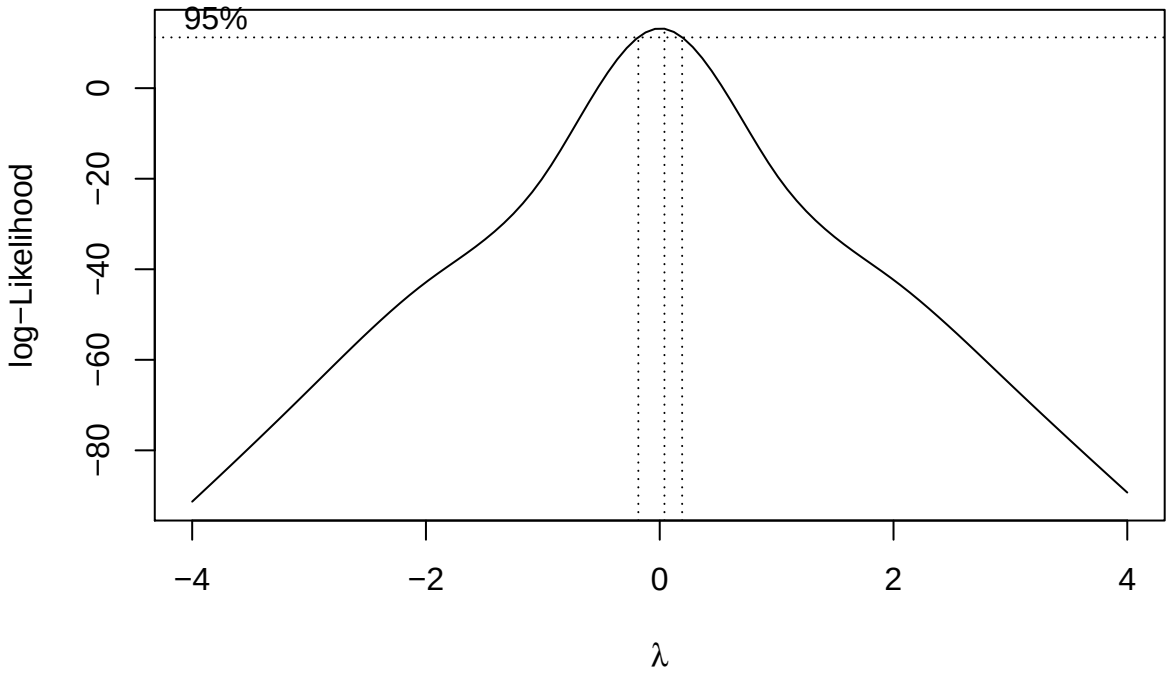
(d.iv.)

A transformation on Y needs to be applied. This is due to the nonconstancy of error variance and the bad normality in error terms.

(e) Box-Cox considers a family of so-called "power transformations", where,

$$Y' = Y^{\lambda}$$

```
library(MASS)
bcmle <- boxcox(concentration.mod,lambda=seq(-4,4),by=0.1)
```



```
lambda <- bcmle$x[which.max(bcmle$y)]
lambda
```

```
## [1] 0.04040404
```

It gives a transfromation that is $Y' = Y^{0.0404}$.

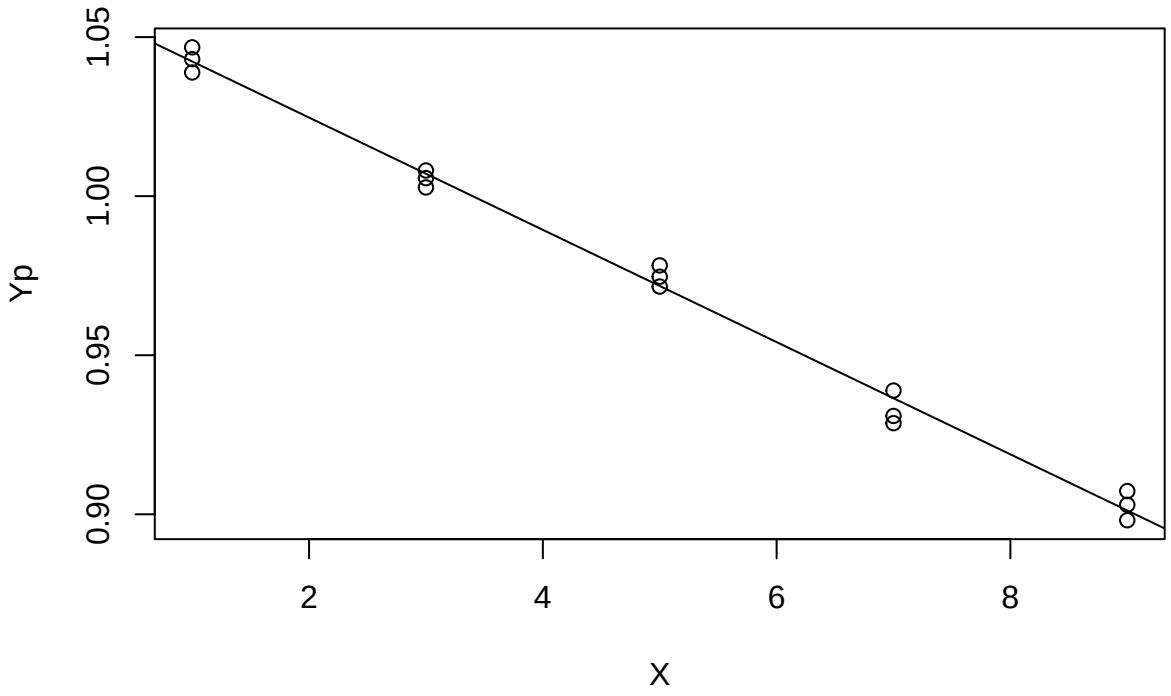
(f) The new model is given by

$$Y' = 1.06 - 0.0176X$$

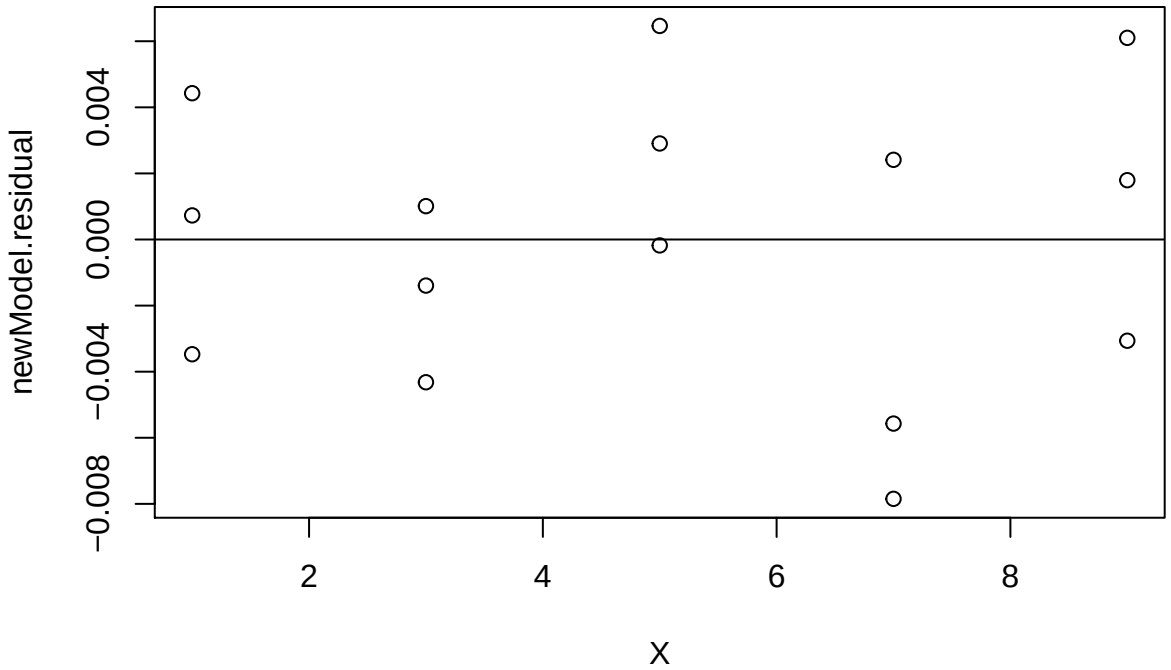
```
Yp <- Y^lambda
newModel <- lm(Yp~X)
summary(newModel)
```

```
##
## Call:
## lm(formula = Yp ~ X)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0078475 -0.0032678  0.0007293  0.0026592  0.0064633
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0599912  0.0023104  458.78  < 2e-16 ***
## X           -0.0176447  0.0004022  -43.87 1.63e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004406 on 13 degrees of freedom
## Multiple R-squared:  0.9933, Adjusted R-squared:  0.9928
## F-statistic:  1925 on 1 and 13 DF,  p-value: 1.627e-15
```

```
plot(X,Yp)
abline(newModel)
```




```
# residual plot
newModel.residual <- residuals(newModel)
plot(X,newModel.residual)
abline(h=0)
```




```
# Brown-Forsythe Test, check the constancy of error variance
library(ALSM)
```

```
## Loading required package: leaps
```

```
## Loading required package: SuppDists
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

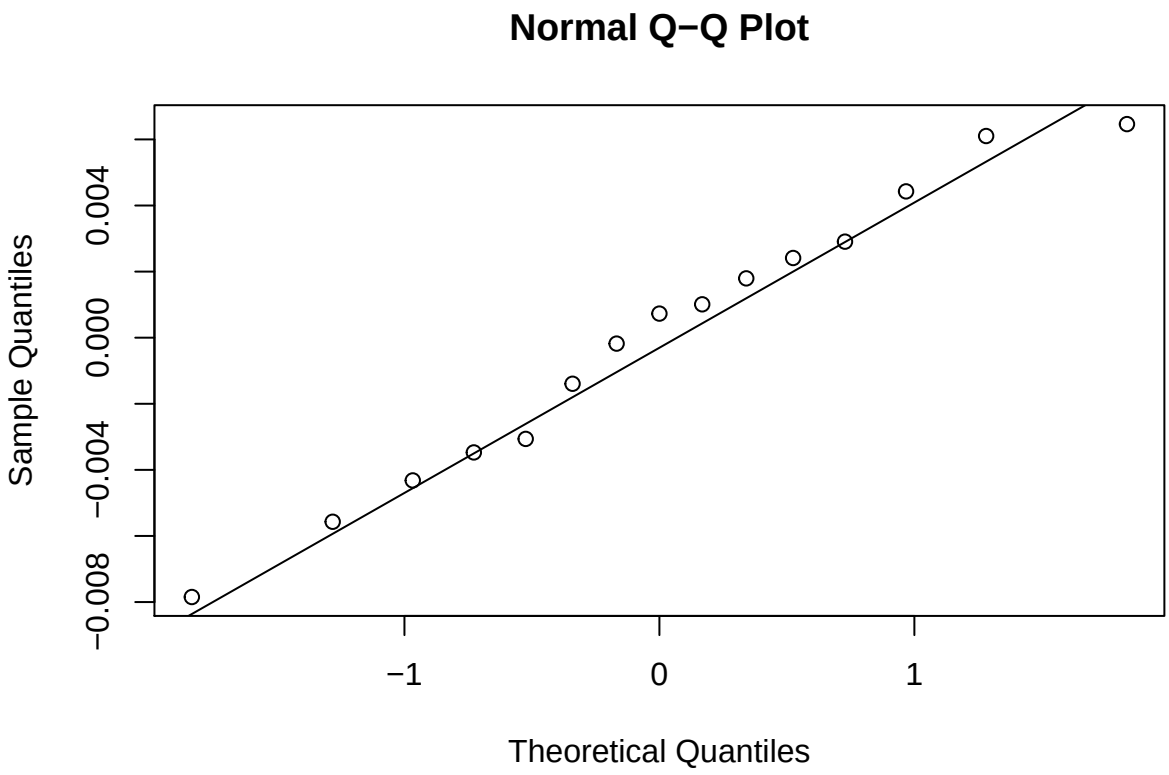
```
g <- rep(1,15)
g[X <= 5] = 0
bftest(newModel,g)
```

```
##       t.value  P.Value alpha df
## [1,] 1.595926 0.134518  0.05 13
```

```
# Sharpiro test, check the normality of the model
shapiro.test(newModel.residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  newModel.residual
## W = 0.97472, p-value = 0.9207
```

```
# probability plot, check the normality of the model
qqnorm(newModel.residual)
qqline(newModel.residual)
```



From these plots and tests, we know that the transformed fitting model is appropriate and the error is randomly normal distributed with constant variance.

(g) Based on the back-transformation, we can compute the confidence inerval for the expected value of Y at X level $X = 7.5$. That is,

```
newModel.cin <- ci.reg(newModel,7.5,type='m',alpha=0.05)
# newModel.cin
oldModel.cin.lower <- (newModel.cin$Lower.Band)^(1 / lambda)
oldModel.cin.upper <- (newModel.cin$Upper.Band)^(1 / lambda)
```

```
cat(paste('The confidence interval of expected value at X = 7.5 is\n [',
          oldModel.cin.lower,',',oldModel.cin.upper,'].'))
```

```
## The confidence interval of expected value at X = 7.5 is
##  [ 0.142809784961862 , 0.170125491521844 ].
```

Interpretation: It means the expected value of Y at a new observation level $X = 7.5$ will lie in this interval with a probability 0.95.

2. If the error terms in a regression model are independent and normally distributed, $N(0, \sigma)$, what can be said about the error terms after transformation $X' = 1/X$ is used? Is the situation the same after transformation $Y' = 1/Y$ is used?

Solution: After transformation $X' = 1/X$, the error terms will still be normally distributed with $N(0, \sigma)$. However, the distribution of error terms will change. From SLM $Y = \beta_0 + \beta_1 X + \epsilon$, predictors are determinant variables, ϵ and Y are random varibales, the variance of Y is same as that of ϵ. A transformaion on Y will modify its distribution and so on for *epsilon*.

3. The following data were obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old.

i	1	2	3	4	5	6	7	8
X	5	8	11	7	13	12	12	6
Y	63	67	74	64	75	69	90	60

(a) Assuming SLM is appropriate, obtain the estimated regression function and plot the residuals against X. What does your residual plot show?
(b) Omit case 7 from the data and obtain the regression function based on the remaining cases. Compare the estimated regression function to (a). What can you conclude about the effect of case 7?
(c) Use your fitted regression in (b), obtain a 99% prediction interval for a new Y observation at $X = 12$.
(d) Use on the data in (b),
   i. Calculate and interpret the 90% individual confidence intervals for β_0 and then for β_1.
   ii. Obtain the Bonferroni joint confidence intervals for β_0 and β_1 using an $\alpha = 0.10$. Interpret your confidence intervals.
   iii. Compare your answers in part i) and ii). Which one of the parts has a larger confidence interval? Why?

Solution:

(a) SLM is estimated below, residuals against X is given. The fitted regression line is given by

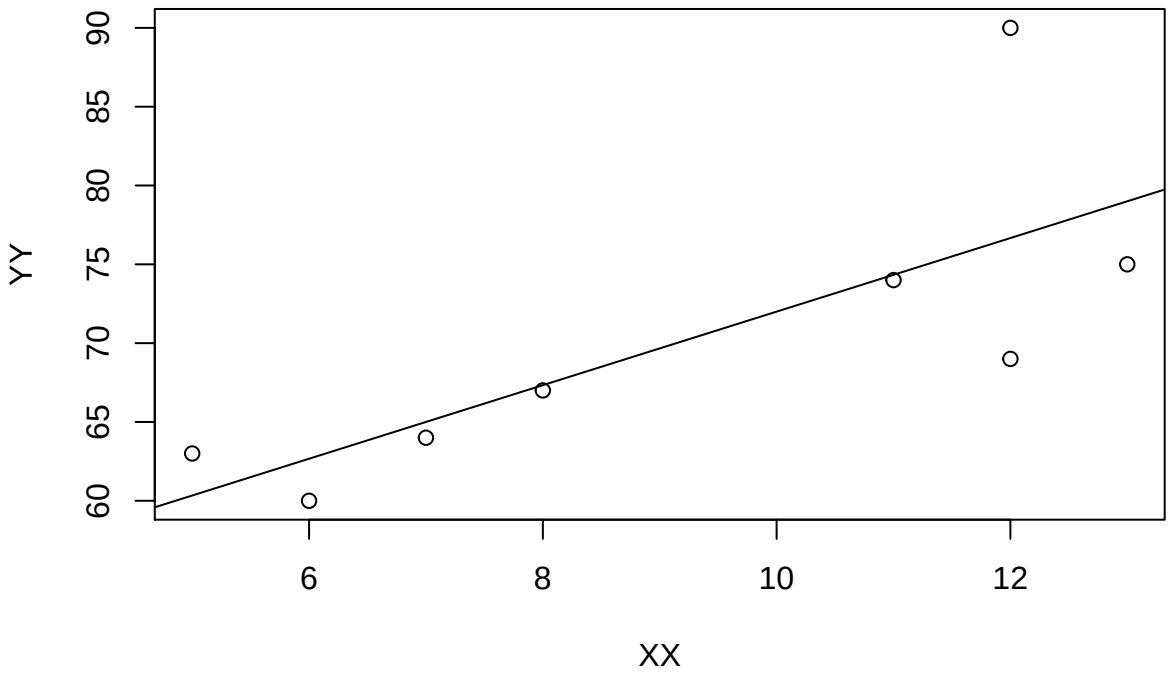
$$Y = 48.6667 + 2.3333X$$

```
XX <- c(5,8,11,7,13,12,12,6)
YY <- c(63,67,74,64,75,69,90,60)
bpag.df <- data.frame(XX,YY)
bpag.mod <- lm(YY~XX,bpag.df)
summary(bpag.mod)
```

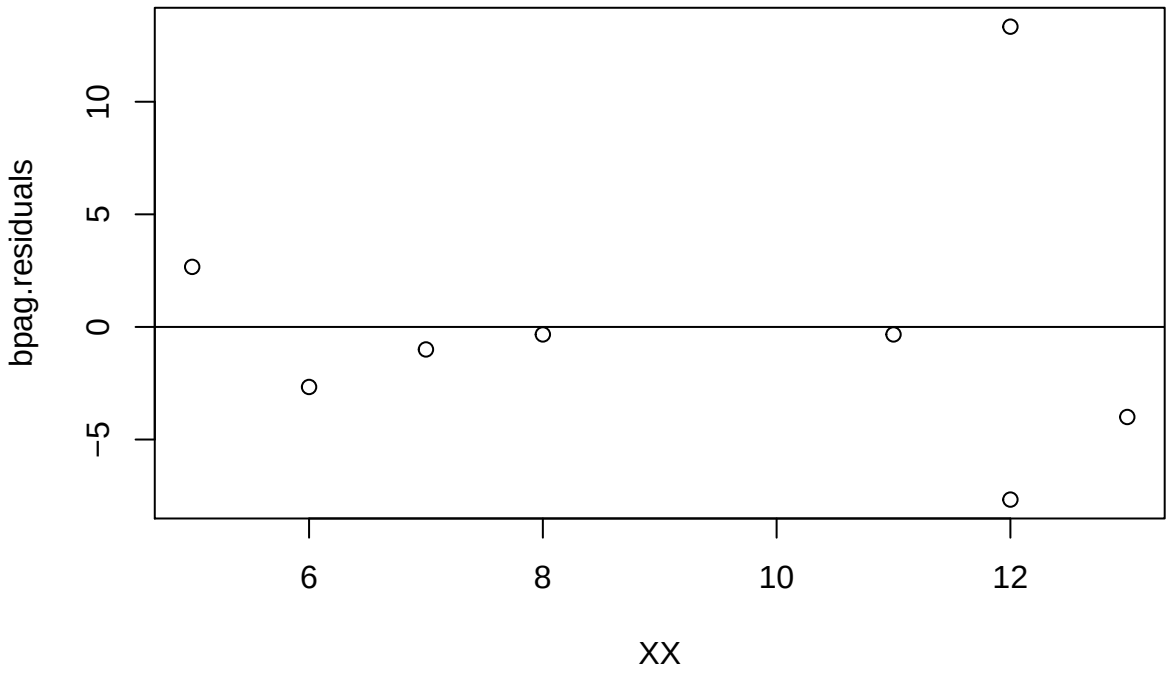
```
##
## Call:
## lm(formula = YY ~ XX, data = bpag.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6667 -3.0000 -0.6667  0.4167 13.3333
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.6667     7.8869   6.171 0.000832 ***
## XX            2.3333     0.8135   2.868 0.028487 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 6 degrees of freedom
## Multiple R-squared:  0.5783, Adjusted R-squared:  0.508
## F-statistic: 8.228 on 1 and 6 DF,  p-value: 0.02849
```

```
plot(XX,YY)
abline(bpag.mod)
```




```
bpag.residuals <- residuals(bpag.mod)
plot(XX,bpag.residuals)
abline(h=0)
```



The residual plot implies that the fitting is not appropriate, the error may not have normal distribution with constant variance. There is outliers at case 7.

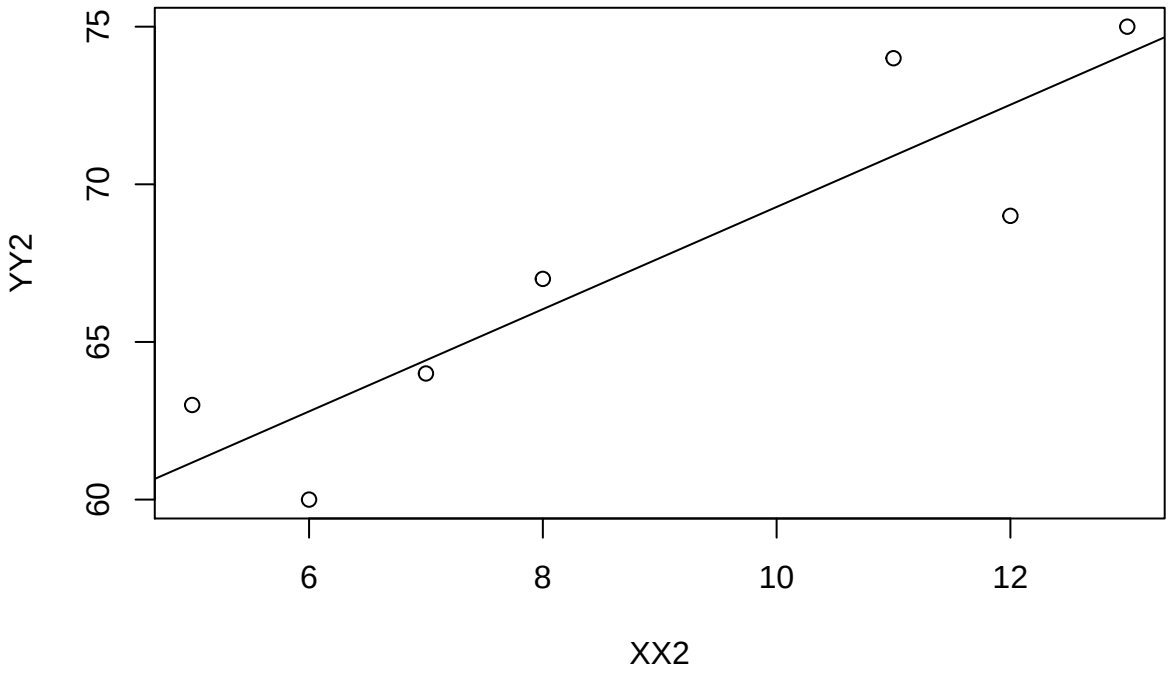
(b) Omit case 7 and we have the linear regression function as

$$Y = 53.068 + 1.6214X$$

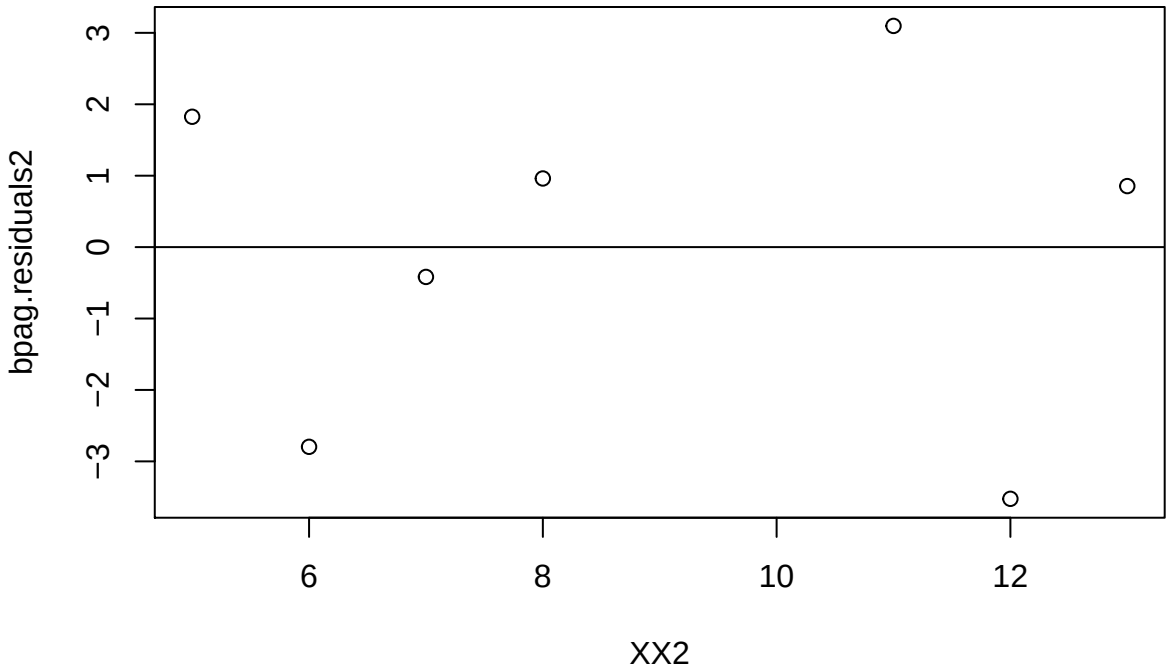
```
XX2 <- c(5,8,11,7,13,12,6)
YY2 <- c(63,67,74,64,75,69,60)
bpag.df2 <- data.frame(XX2,YY2)
bpag.mod2 <- lm(YY2~XX2,bpag.df2)
summary(bpag.mod2)
```

```
##
## Call:
## lm(formula = YY2 ~ XX2, data = bpag.df2)
##
## Residuals:
##       1       2       3       4       5       6       7
##  1.8252  0.9612  3.0971 -0.4175  0.8544 -3.5243 -2.7961
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.0680     3.2136  16.514 1.49e-05 ***
## XX2           1.6214     0.3448   4.702  0.00533 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.645 on 5 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7787
## F-statistic: 22.11 on 1 and 5 DF,  p-value: 0.005327
```

```
plot(XX2,YY2)
abline(bpag.mod2)
```




```
bpag.residuals2 <- residuals(bpag.mod2)
plot(XX2,bpag.residuals2)
abline(h=0)
```



Comparing the linear regression function in (b) to (a) implies that the case 7 is an outlier point which can make the estimated variance of error larger than its true value.

(c) The prediction confidence interval for a new observation at $X = 12$ is given by

```
bpag.ci <- ci.reg(bpag.mod2,12,type='n',alpha=0.01)
cat(paste('The prediction confidence interval at X = 12 is\n [',
          bpag.ci$Lower.Band,',',bpag.ci$Upper.Band,'].'))
```

```
## The prediction confidence interval at X = 12 is
##  [ 60.3126605269493 , 84.7358831623711 ].
```

(d) i. The individual confidence interval for β_0 and β_1 is given below.

```
confint(bpag.mod2,level=0.9)
```

```
##                    5 %      95 %
## (Intercept) 46.5923996 59.543523
## XX2          0.9265355  2.316183
```

Confidence interval for β_0 is $[46.59, 59.54]$. Confidence interval for β_1 is $[0.9265, 2.3162]$. This maens the true value of β_0 and β_1 will fall into these intervals with probability 0.9.

ii. The Bonferroni joint confidence interval for β_0 and β_1 are given by

```
confint(bpag.mod2,level=0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept) 44.8071367 61.328786
## XX2          0.7349778  2.507741
```

The true value of β_0 will fall into $[44.8071, 61.3288]$ and the true value of β_1 will fall into $[0.735, 2.508]$ with joint probability 0.9.

iii. The Bonferroni joint intervals for β_0 and β_1 will be wider than the individual confidence intervals. This is because the $1-\alpha$ joint (family) confidence interval is done by estimating β_0 and β_1 separately with individual confidence level of $1-\alpha/2$ each.

4. Based on the following small data set, construct by hand the design matrix, X, its transpose X', and the matrices $X'X$, $(X'X)^{-1}$, $X'Y$, $b = (X'X)^{-1}X'Y$, the variance-covariance matrix of $\Sigma\{b\}$ and $\Sigma\{\hat{Y}\}$ (i.e., $s^2\{b\}$ and $s^2\{\hat{Y}_h\}$ in matrix form).

X	Y
2	1
3	2
6	4
7	5
9	7

Solution: The matrix forms are given by

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 7 \\ 1 & 9 \end{bmatrix}, \quad X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 6 & 7 & 9 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 5 & 27 \\ 27 & 179 \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} 1.0783133 & -0.16265060 \\ -0.1626506 & 0.03012048 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 19 \\ 130 \end{bmatrix}, \quad b = (X'X)^{-1}X'Y = \begin{bmatrix} -0.6566265 \\ 0.8253012 \end{bmatrix}$$

$$\Sigma\{b\} = MSE \cdot (X'X)^{-1} = 0.062249 \cdot \begin{bmatrix} 1.0783133 & -0.16265060 \\ -0.1626506 & 0.03012048 \end{bmatrix}$$
$$= \begin{bmatrix} 0.06712392 & -0.01012484 \\ -0.01012484 & 0.00187497 \end{bmatrix}$$

$$\Sigma\{\hat{Y}\} = MSE \cdot \left(X_h'(X'X)^{-1}X_h\right) = X_h'\Sigma\{b\}X_h$$

where X_h is a matrix in form $X_h = \begin{bmatrix} 1 \\ x_h \end{bmatrix}$, x_h is the new predictor at which the mean response is solved.

```
XXX <- matrix(c(1,1,1,1,1,2,3,6,7,9),nrow = 5, ncol = 2,byrow = FALSE)
t(XXX)%*%XXX
```

```
##      [,1] [,2]
## [1,]    5   27
## [2,]   27  179
```

```
xtxinv <- solve(t(XXX)%*%XXX)
xtxinv
```

```
##            [,1]        [,2]
## [1,]  1.0783133 -0.16265060
## [2,] -0.1626506  0.03012048
```

```
YYY <- c(1,2,4,5,7)
t(XXX)%*%YYY
```

```
##      [,1]
## [1,]   19
## [2,]  130
```

```
b <- solve(t(XXX)%*%XXX)%*%(t(XXX)%*%YYY)
b
```

```
##            [,1]
## [1,] -0.6566265
## [2,]  0.8253012
```

```
YYYhat <- XXX %*% b
YYYhat
```

```
##           [,1]
## [1,] 0.9939759
## [2,] 1.8192771
## [3,] 4.2951807
## [4,] 5.1204819
## [5,] 6.7710843
```

```
SSE <- t(YYY - YYYhat) %*% (YYY - YYYhat)
MSE <- SSE / (5-2)
MSE
```

```
##          [,1]
## [1,] 0.062249
```

```
drop(MSE) * xtxinv
```

```
##             [,1]        [,2]
## [1,]  0.06712392 -0.01012484
## [2,] -0.01012484  0.00187497
```

5. When joint confidence intervals for β_0 and β_1 are developed by the Bonferroni method with a family confidence coefficient of 90 percent, does this imply that 10 percent of the time the confidence interval for β_0 will be incorrect? That 5 percent of the time the confidence interval for β_0 will be incorrect and 5 percent of the time that for β_1 will be incorrect? Discuss.

Solution: The Bonferroni method will imply that 5 percent of the time the confidence interval for β_0 will be incorrect and 5 percent of the time that for β_1 will be incorrect. If we assume event I_0^+ means the I_0 successfully contain the true intercept β_0 and I_1^+ is the event that I_1 successfully contain the true slope β_1. $P(I_0^+$ and $I_1^+) \geq 1 - \alpha$ means that we require $P(I_0^+) = 1 - \alpha/2$ and $P(I_1^+) = 1 - \alpha/2$ respectively.

6. A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

Solution: The residual aginst the fitted Y ($\hat{Y}_i$) will be a more meaningful plot. Since we have $Y_i = \hat{Y}_i + e_i$, fitted Y and residuals e_i are independent, Y_i will have a positive relations with e_i especially when the fitted line is flat, which indicated this relationship plot is that meaningful for our analysis.