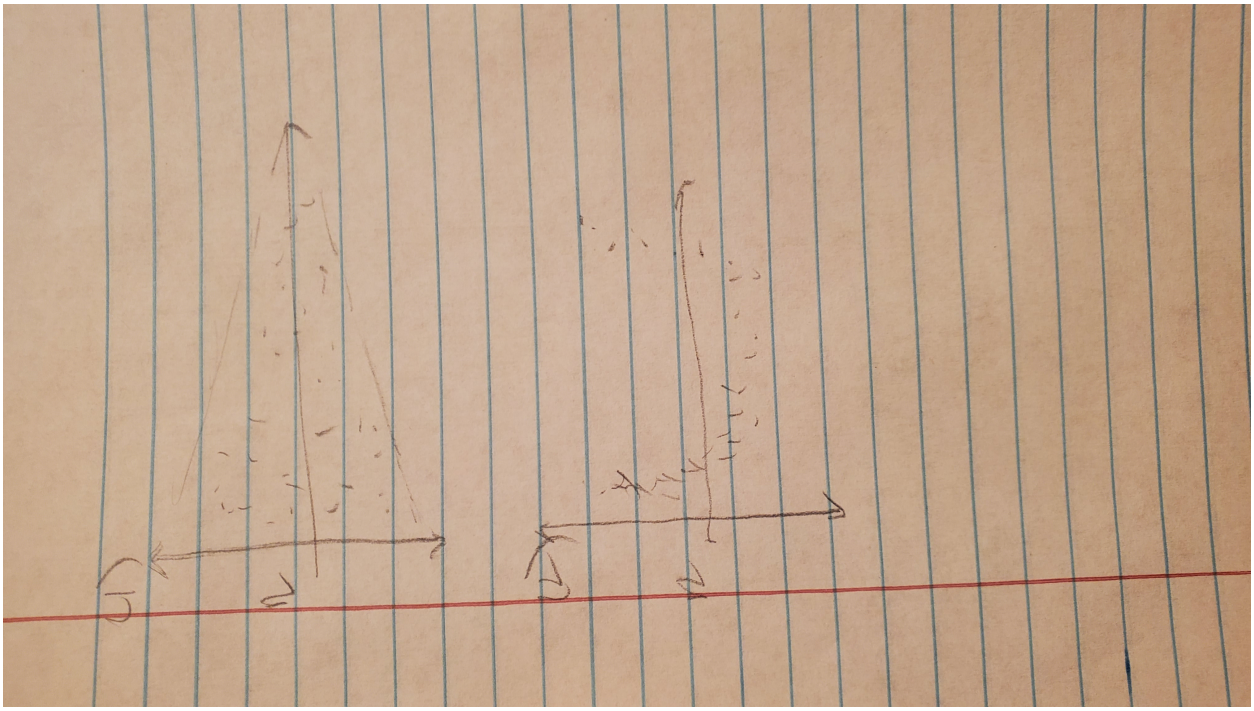
Homework 4

Ishita Dutta

4/30/2021

3.2

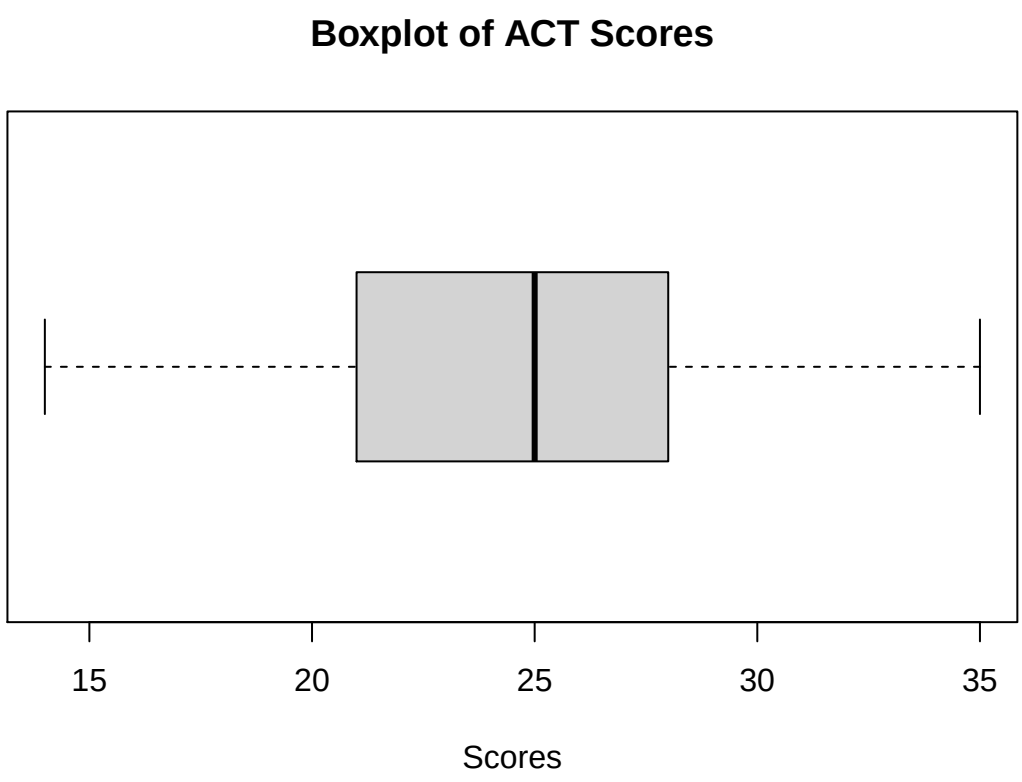
```
knitr::include_graphics("20210501_000812.jpg")
```



3.3

a)

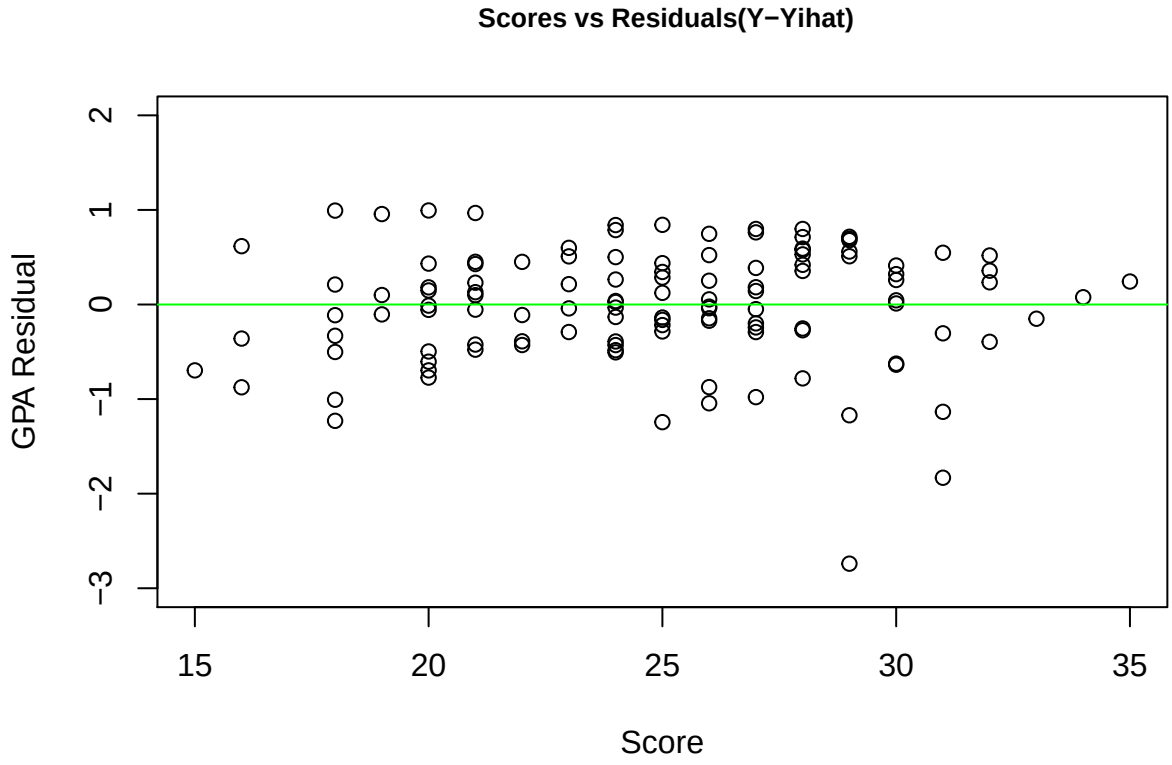
```
GPAdata = read.table("grade+point+average.txt")
colnames(GPAdata) <- c("y", "x", "x2", "x3")
X1 = GPAdata[,2]
numvalsx = length(X1)
boxplot(X1, horizontal = TRUE, xlab = "Scores")
title("Boxplot of ACT Scores")
```



The thing we can note is that the distribution is relatively symmetric.

b)

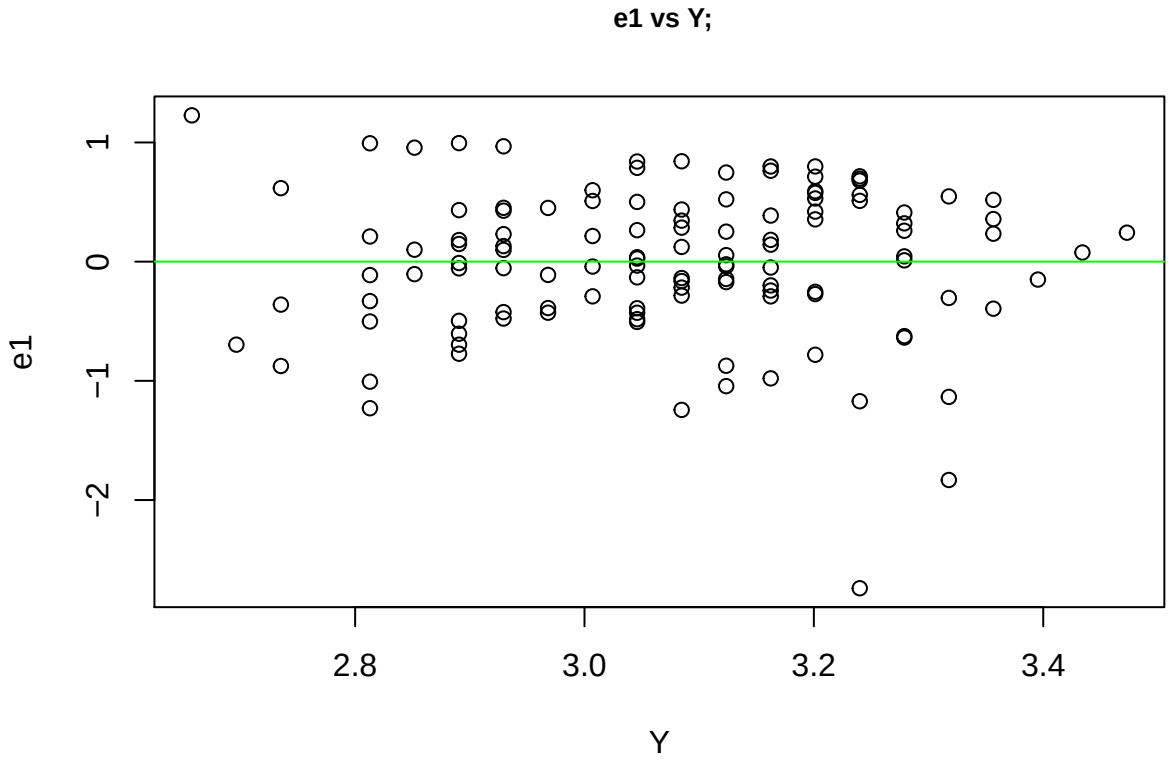
```
Y1 = GPAdata[,1]
numVals = length(Y1)
fit1 = lm(Y1~X1)
bohat1 = fit1$coefficients[[1]]
b1hat1 = fit1$coefficients[[2]]
Yi_hat1 = bohat1 + b1hat1 * (X1)
Y1_1 = Y1-Yi_hat1
fit1 = lm(Y1_1~X1)
plot(X1,Y1_1,
 xlim = c(15,35), ylim=c(-3,2),
 main = " Scores vs Residuals(Y-Yihat)",
 xlab = "Score",
 ylab = "GPA Residual",
 cex.main = .85)
abline(fit1, col = "green")
```



Note: values are relatively close to regression line save for the one point that seems to have a GPA residual of -3 but an ACT of 29.

c)

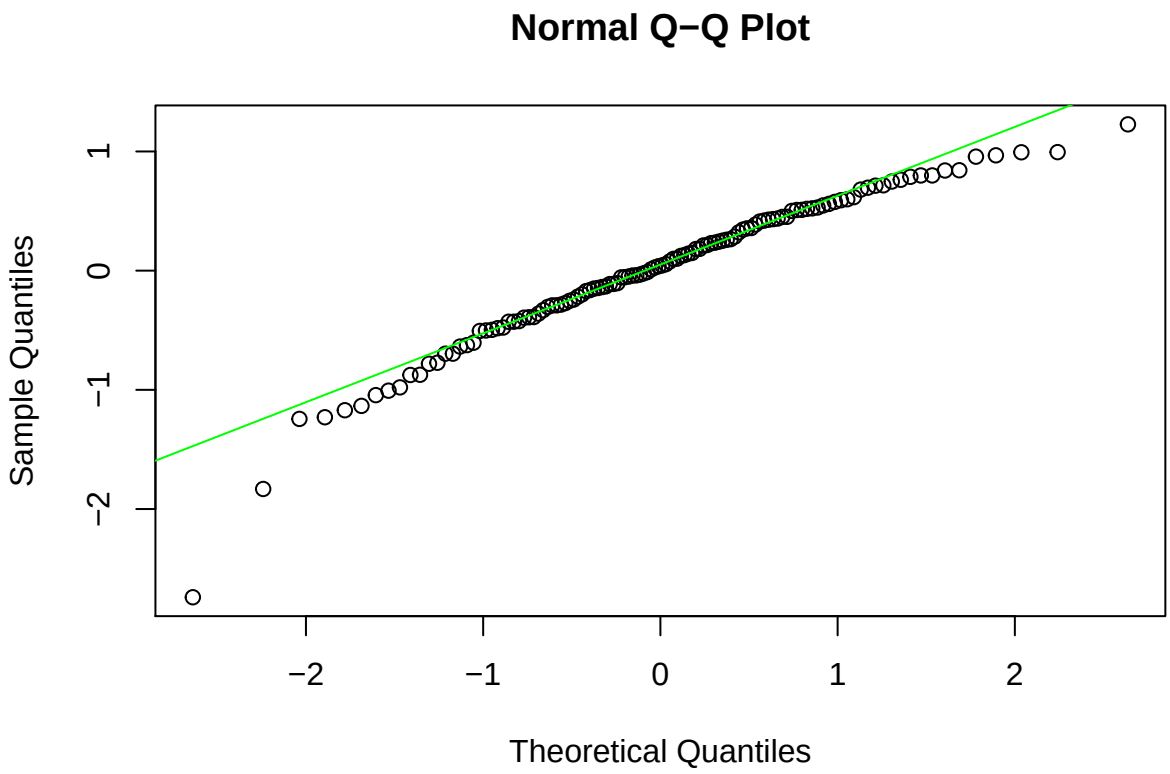
```
e1 = fit1$residuals
fit1_1 = lm(e1~Yi_hat1)
plot(Yi_hat1,e1,
 main = " e1 vs Y;",
 xlab = "Y",
 ylab = "e1",
 cex.main = .85)
abline(fit1_1, col = "green")
```



From this model, we can study the distribution of B1 over the GPA values, illustrated by e1.

d)

```
res = qqnorm(e1)
qqline(e1, col='green')
```




```
r=cor(res$x, res$y)#res$x is expected residuals, res$y is observed residuals
cat("r = ", r)
```

```
## r =  0.9744497
```

With alpha = 0.05, the critical value at n = 120 is 0.987, which is greater than r = 0.974, meaning that the distribution is not normal

e)

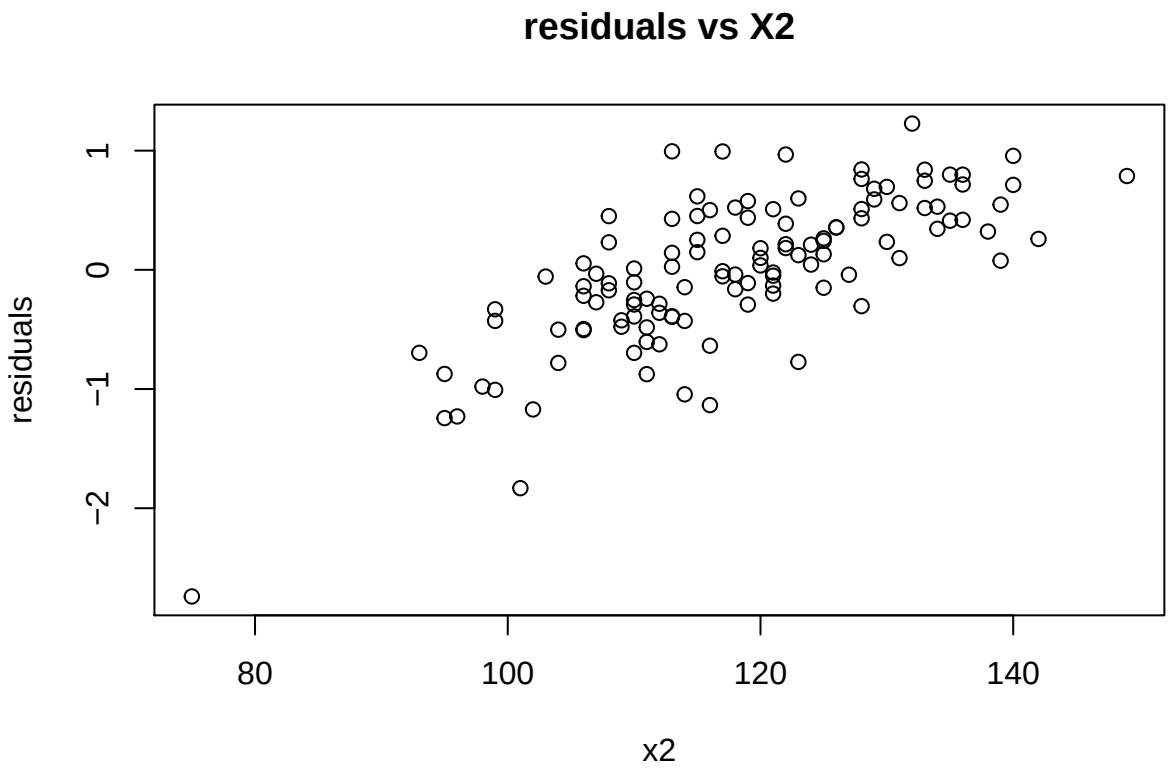
```
group1 <- GPAdata$x < 26
group2 <- !group1
d1 <- abs(e1[group1] - median(e1[group1]))
d2 <- abs(e1[group2] - median(e1[group2]))
n <- length(e1)
n1 <- length(d1)
n2 <- length(d2)
s <- sqrt((sum((d1-mean(d1))^2)+sum((d2-mean(d2))^2))/(n - 2))
tstar <- (mean(d1)-mean(d2))/s/sqrt(1/n1+1/n2)
tcrit <- qt(0.995, df = n-2)
abs(tstar) <= tcrit
```

```
## [1] TRUE
```

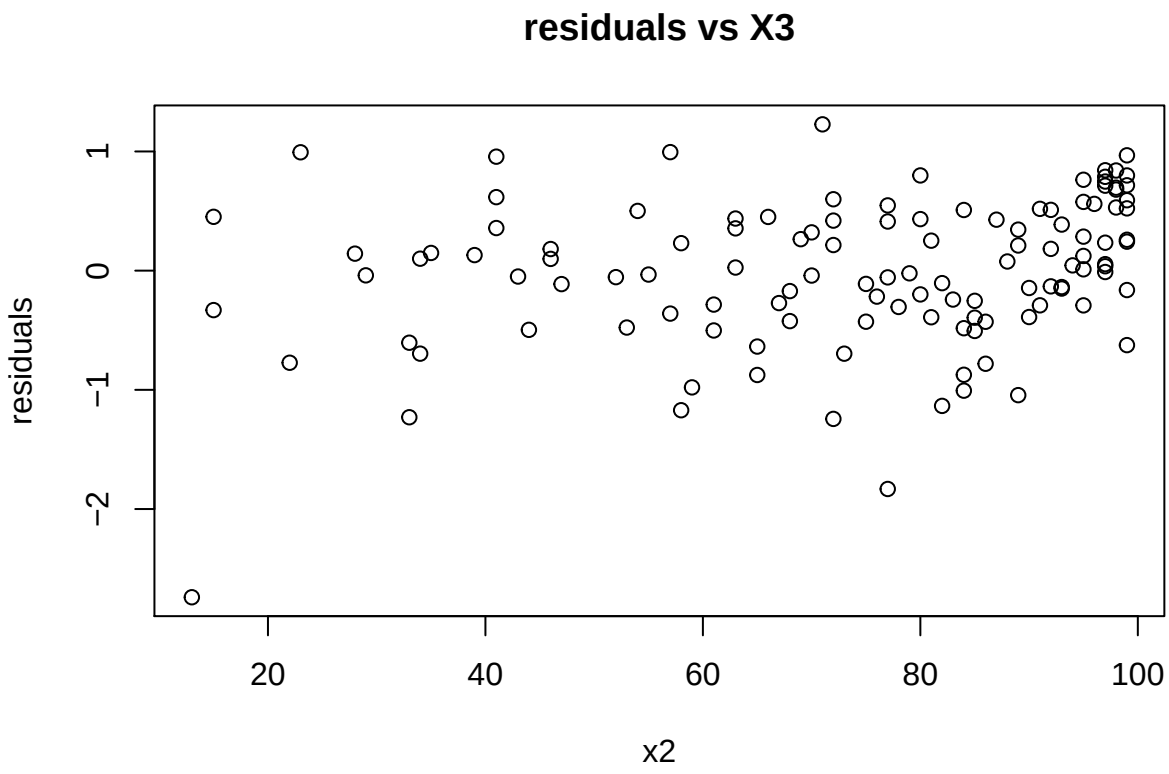
Because |tstar| < tcrit, we conclude the null(Ho), where the error variance is constant as opposed to the alternate(Ha) of the error variance not being constant

f)

```
plot(GPAdata[,3], e1, xlab = "x2", ylab = "residuals", main = "residuals vs X2")
```




```
plot(GPAdata[,4], e1, xlab = "x2", ylab = "residuals", main = "residuals vs X3")
```

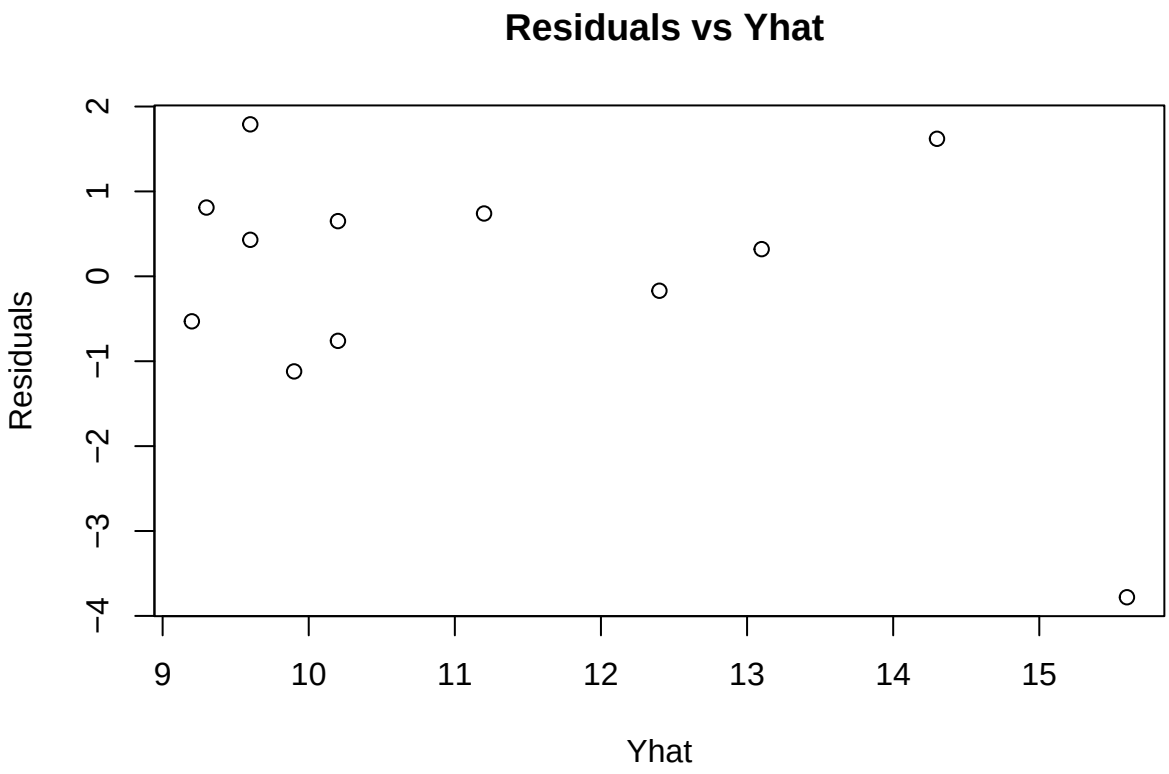


The X2 plot shows a nonconstant relationship, so we should consider it as a regressor. On the other hand, X3 shows a constant relationship, so we should not be missing much if we omit it as a regressor.

3.10

a)

```
PCEdata <- read.table("per+capita+earnings.txt")
yhat2 = PCEdata[,1]
e2 = PCEdata[,2]
fit2 = lm(e2~yhat2)
plot(yhat2,e2,
 main = " Residuals vs Yhat",
 xlab = "Yhat",
 ylab = "Residuals")
```



The p1ot suggests values generally above the proposed regression line. The reason might be from the possible outlier at a residual of -4.

##b)

```
# Given our sd = 1
sum(e2 > 1 | e2 < -1)
```

```
## [1] 4
```

We have 4 values outside 1 standard deviation. Assuming the normal distribution, we would expect to see 12 x P(|Z| > 1) = 3.81, so approximately 4 as well.

3.13

a

b

c

3.15

a

```
Soldata <- read.table("solution+concentration.txt")
X4 = Soldata[,2]
Y4 = Soldata[,1]
b1hat4 = t(X4-mean(X4))%*%(Y4-mean(Y4))/sum((X4-mean(X4))^2)
b0hat4 = mean(Y4) - b1hat4*mean(X4)
cat("Y = ", b0hat4, " + ", b1hat4, "x\n")
```

```
## Y =  2.575333  +  -0.324 x
```

b

Ho = E(Y) = B0 + B1 * X Ha = E(Y) != B0 + B1 * X SSPE = 0.1575 MSPE = 0.1575/10 = 0.0157 SSE = 2.9247 SSLF = SSE - SSPE = 2.7675 MSLF = 2.7675/3 = 0.9224 F* =58.5714 F = 4.825621

c

```
knitr::include_graphics("20210430_235736.jpg")
```