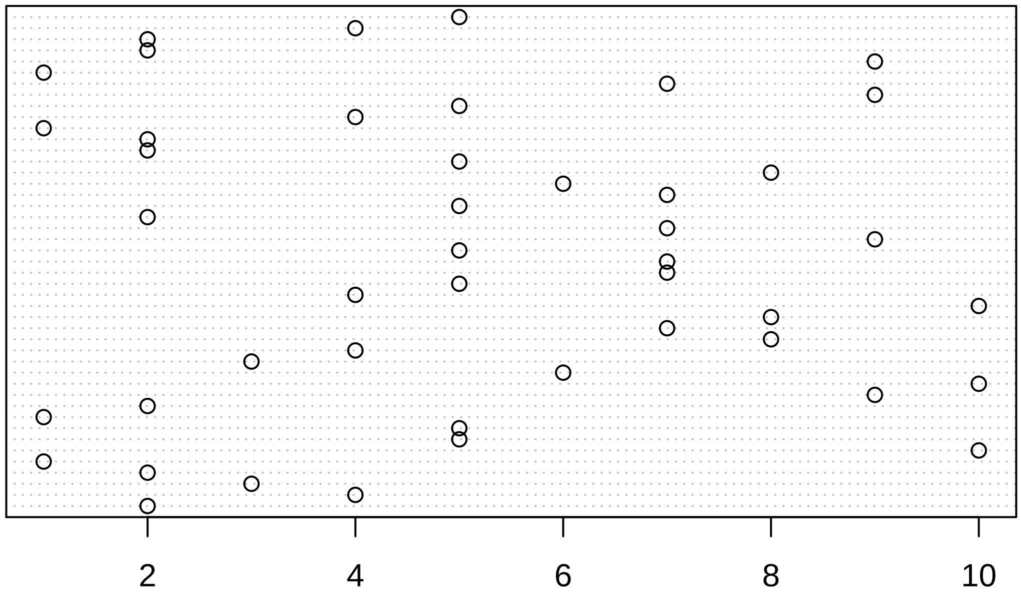
STAT GR5205 – Section 005 HW 3

Bo Rong br2498

Oct. 10th, 2016

```
#1.
#(a)
filename <- "~/Downloads/copiers_full.txt"
copiers_full<- read.table(file=filename, header=T)
x <- copiers_full$copiers
y <- copiers_full$minutes
dotchart(x)
```

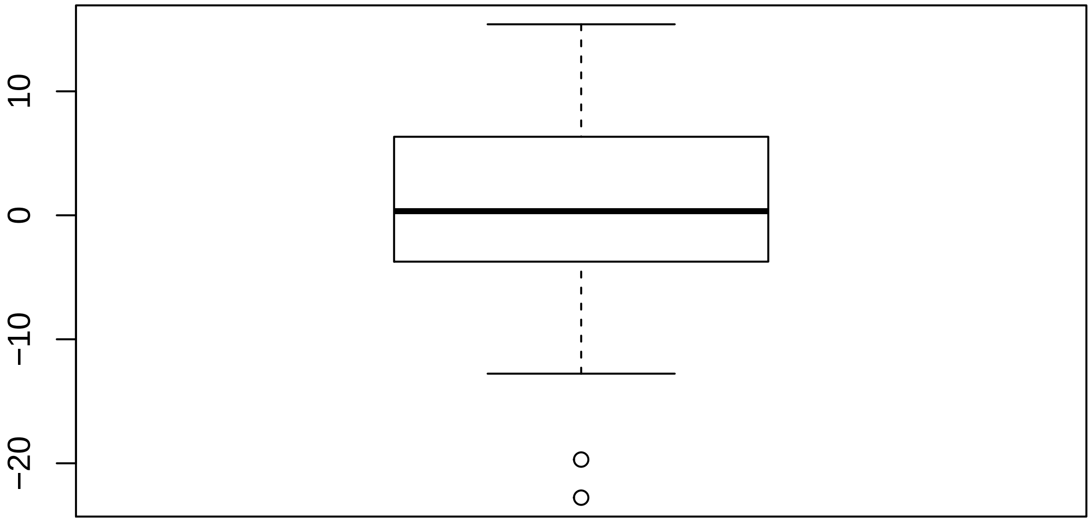


```
#The plot is regular,every value is represented. There is no outliers and there is no indication of temp
```

```
#(b)
fit<-lm(y~x)
summary(fit)
```

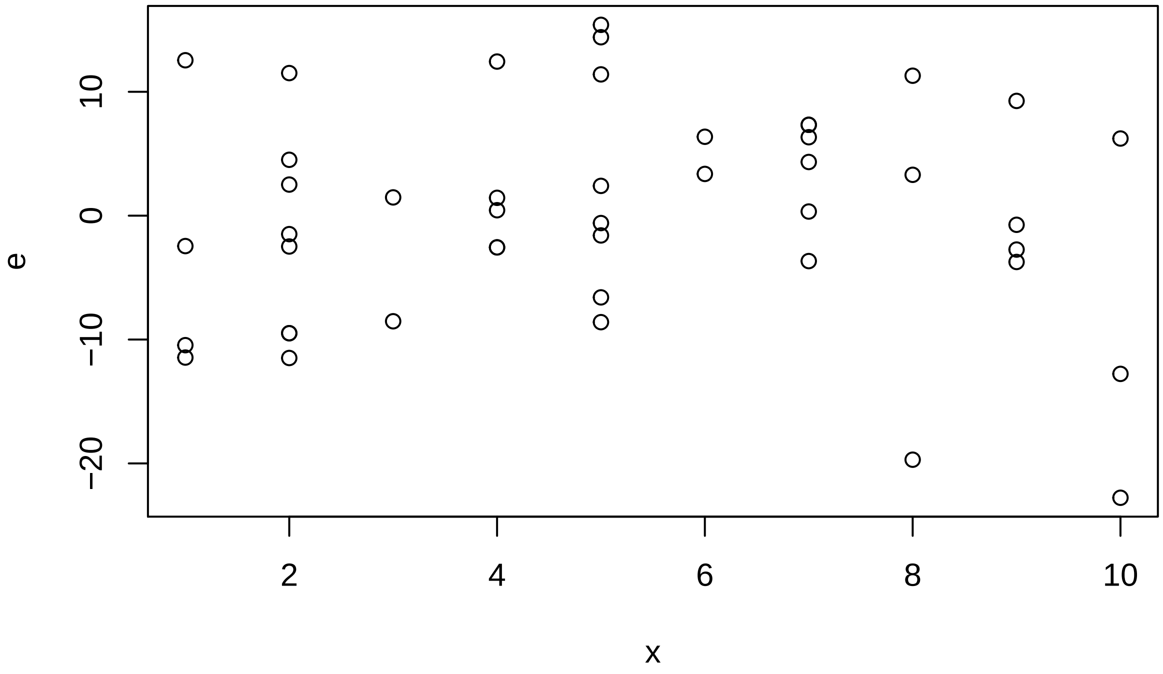
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## x            15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

```
boxplot(resid(fit))
```

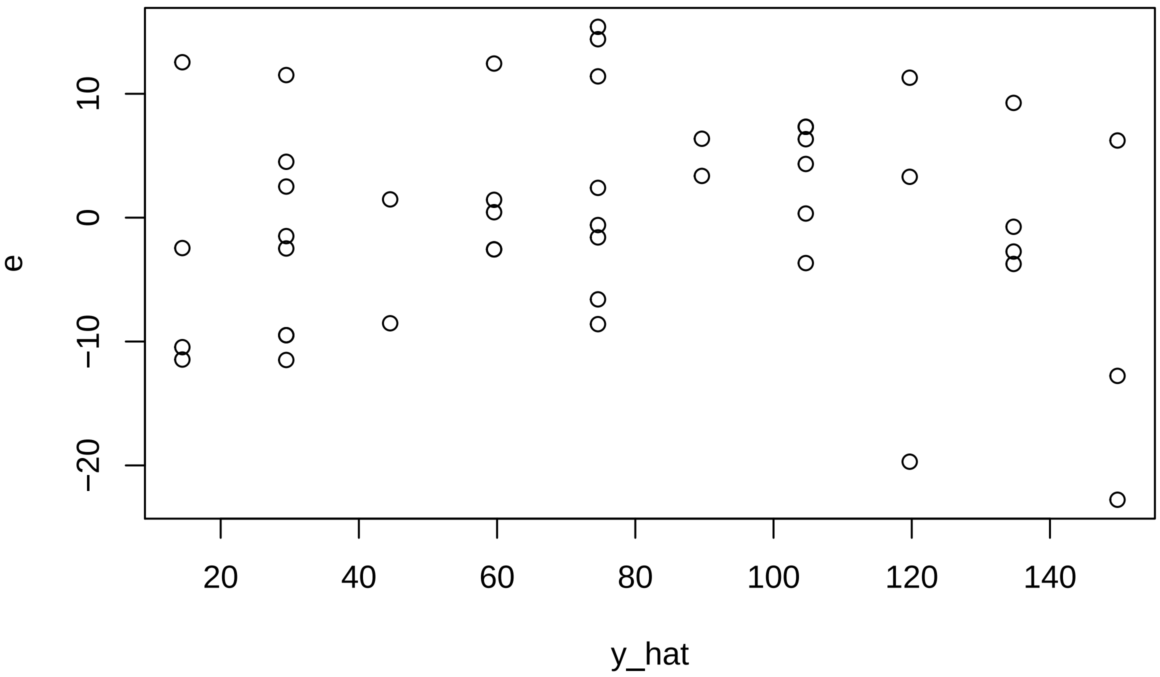


```
#Not completely symmetric. The distribution of residuals is slightly skewed to the positive side and ha
```

```
#(c)
e<-resid(fit)
plot(x,e)
```

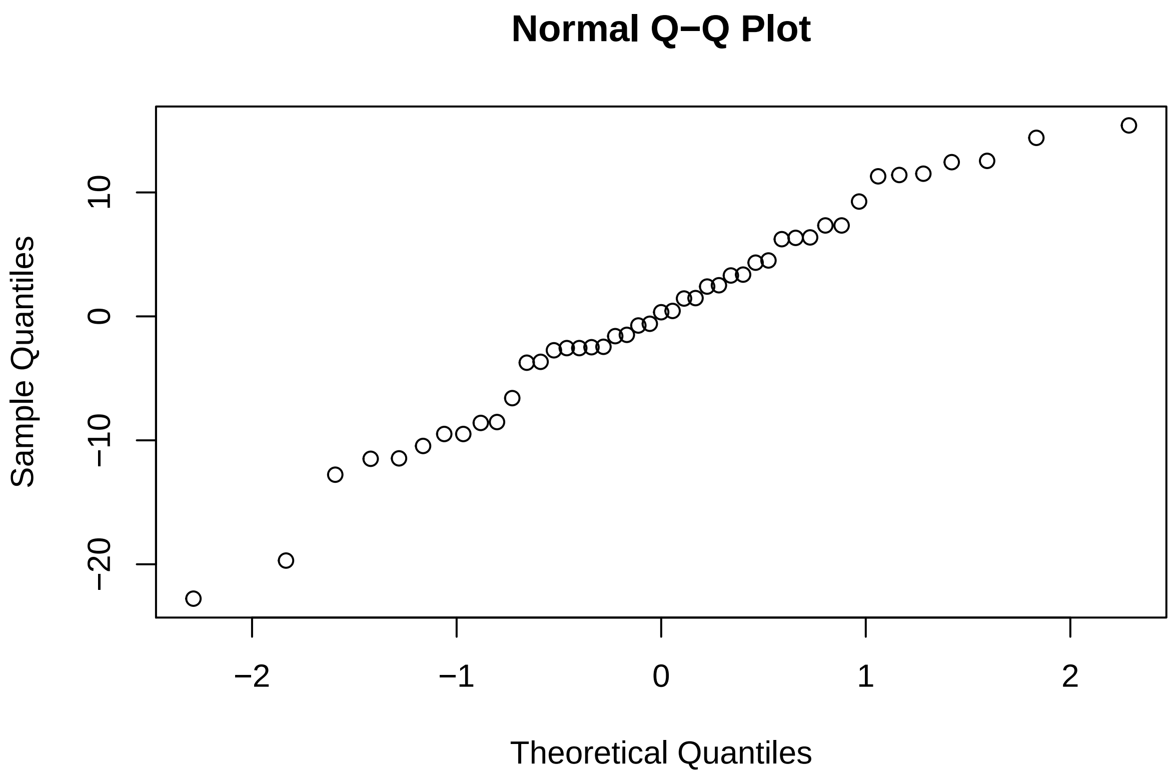


```
y_hat<-fitted(fit)
plot(y_hat,e)
```



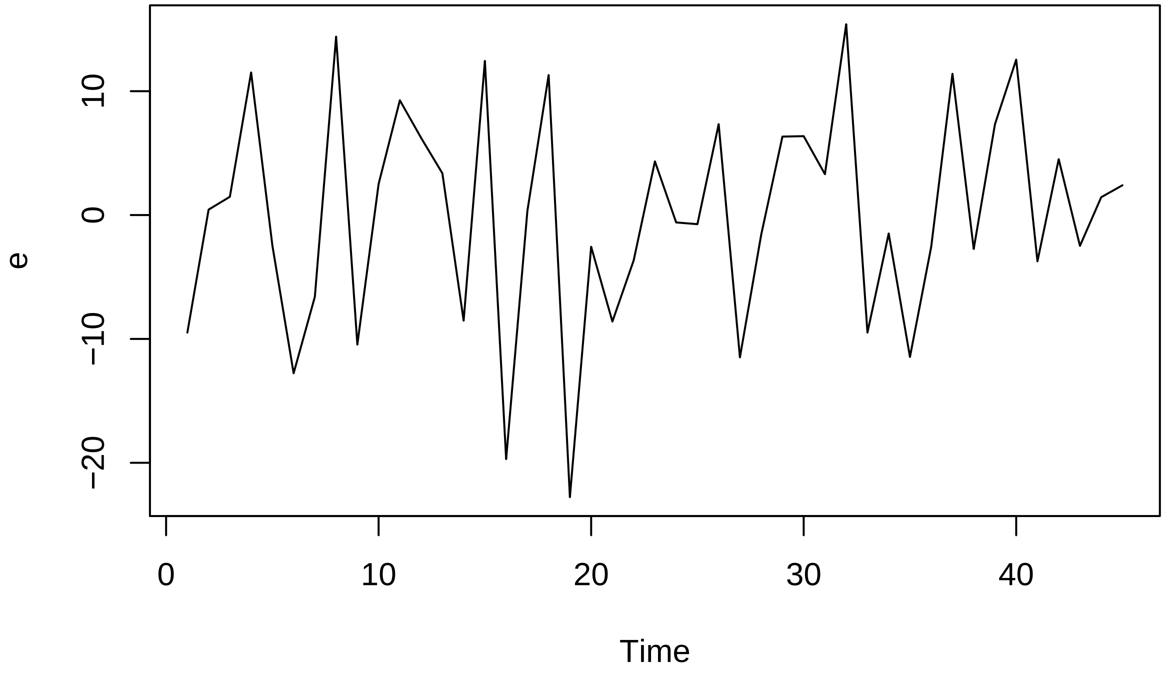
```
#These two plots provide the same information.Because the fitted values and predictor variable
#are linearly related. These plots indicate no severe departure from the assumption of constant varianc
```

```
#(d)
qqnorm(e)
```



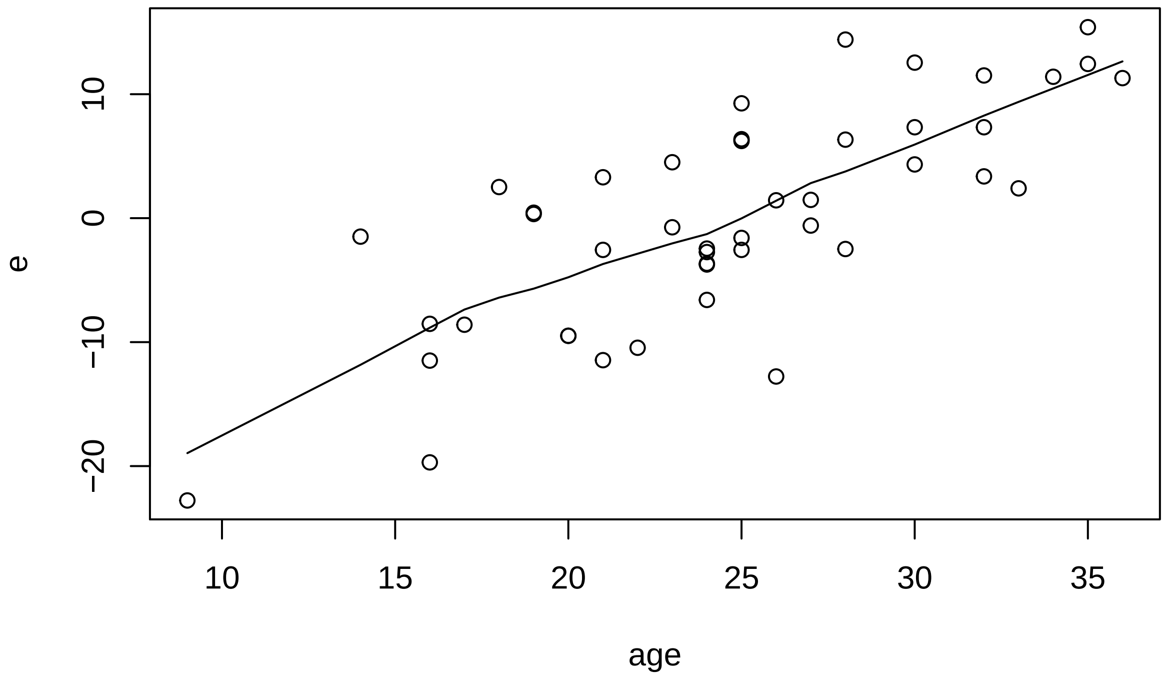
```
#The plot seems a straight line if we don't consider that two points.
```

```
#(e)
plot.ts(e)
```

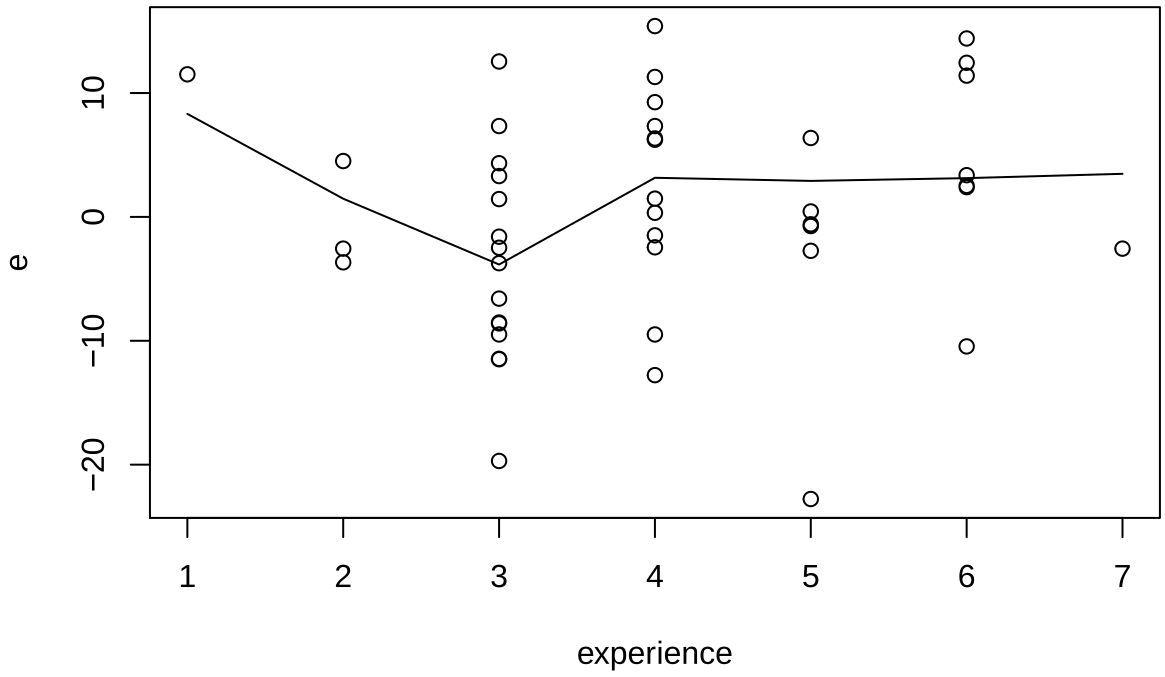


```
#Error terms are not correlated over time, because there is no tendency for residuals to appear positiv
```

```
#(f)
plot(e ~ age, data=copiers_full)
lines(lowess(e ~ copiers_full$age))
```

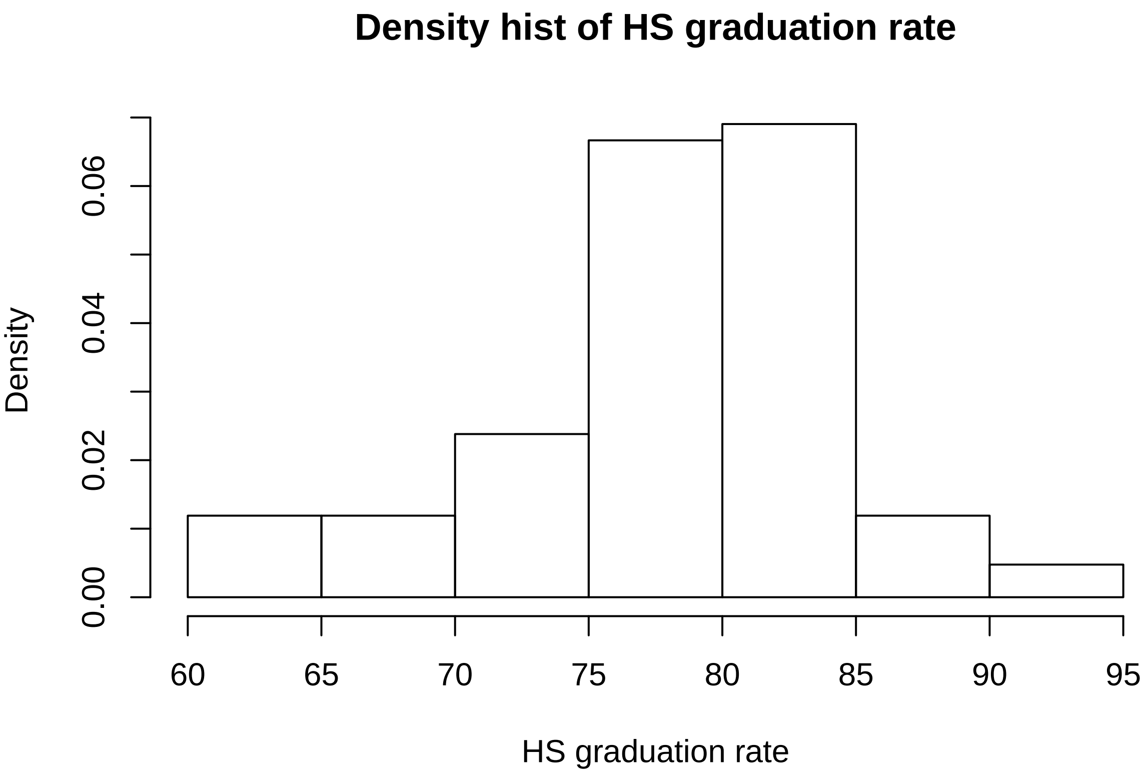


```
plot(e ~ experience, data=copiers_full)
lines(lowess(e ~ copiers_full$experience))
```



```
#The age of the copiers gets larger, the person need more time to fix copiers.
#The experience does not have any additional effect for predicting service time.
```

```
#2
#(a)
filename <- "~/Downloads/crime_rates.txt"
crime_rates <- read.table(file=filename, header=T)
hist(crime_rates$x,freq=F,xlab="HS graduation rate",main = "Density hist of HS graduation rate")
```

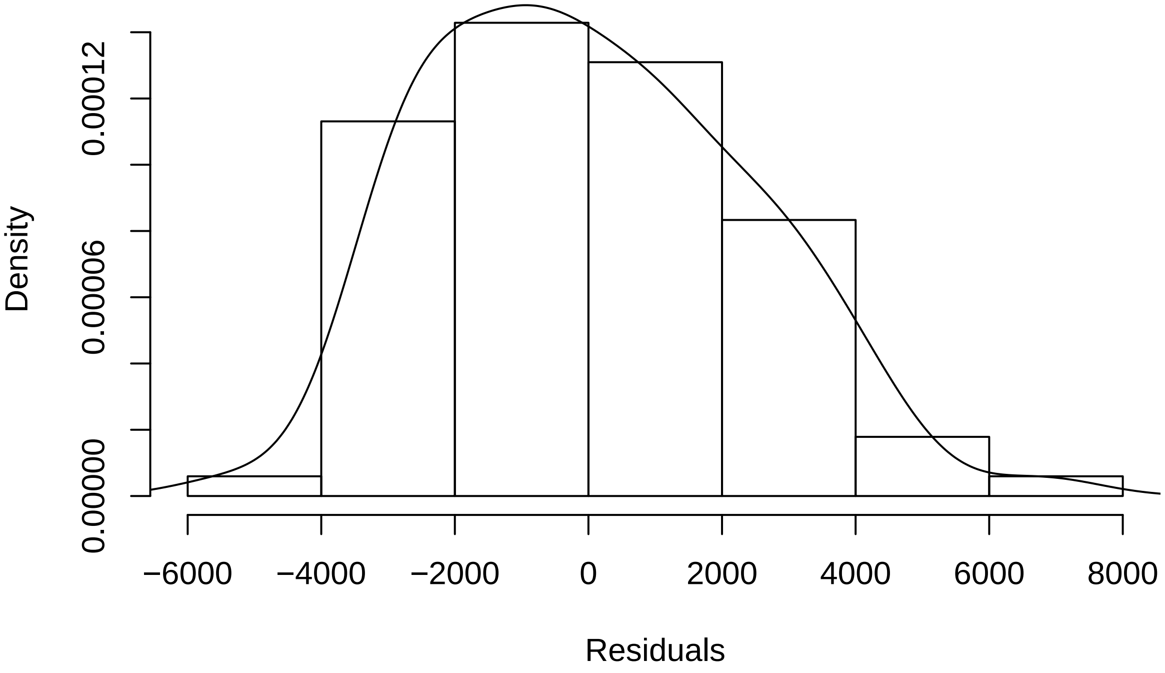


```
#Distribution is skewed to the left. The HS graduation rate of all counties is between 60 and 95.
#And most of counties has HS graduation rate between 75 and 85.
```

```
#(b)
fit <- lm(y ~ x, data=crime_rates)
summary(fit)
```

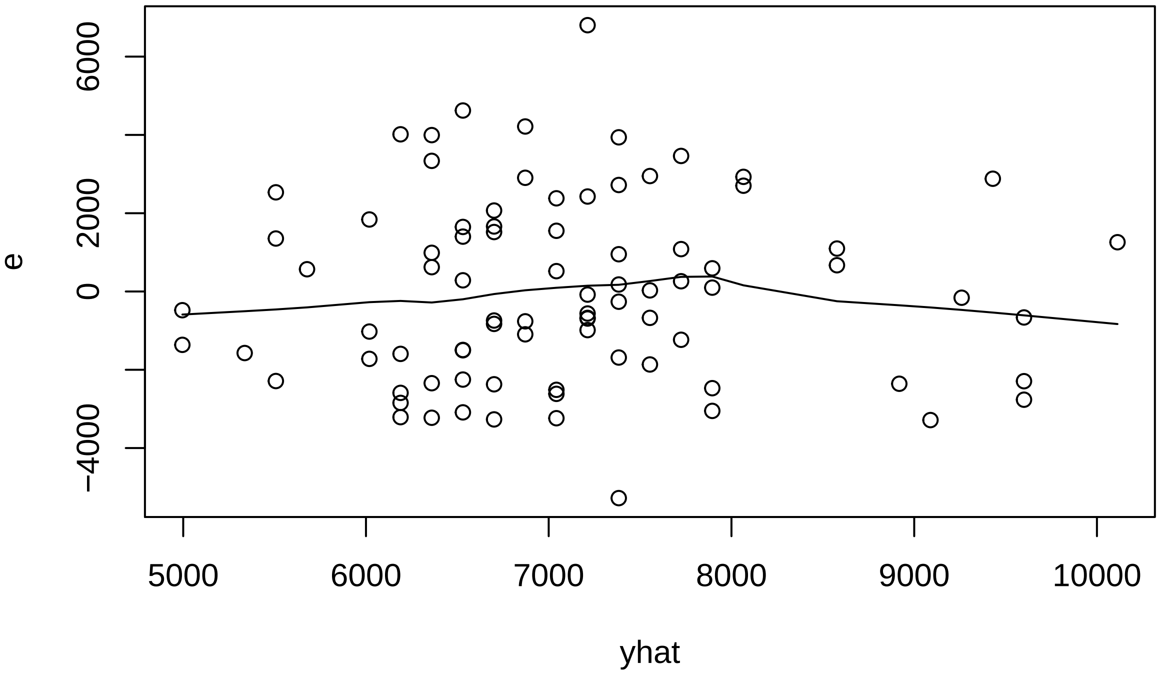
```
##
## Call:
## lm(formula = y ~ x, data = crime_rates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## x             -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

```
hist(resid(fit), freq=F,xlab="Residuals", main="")
lines(density(resid(fit)))
```



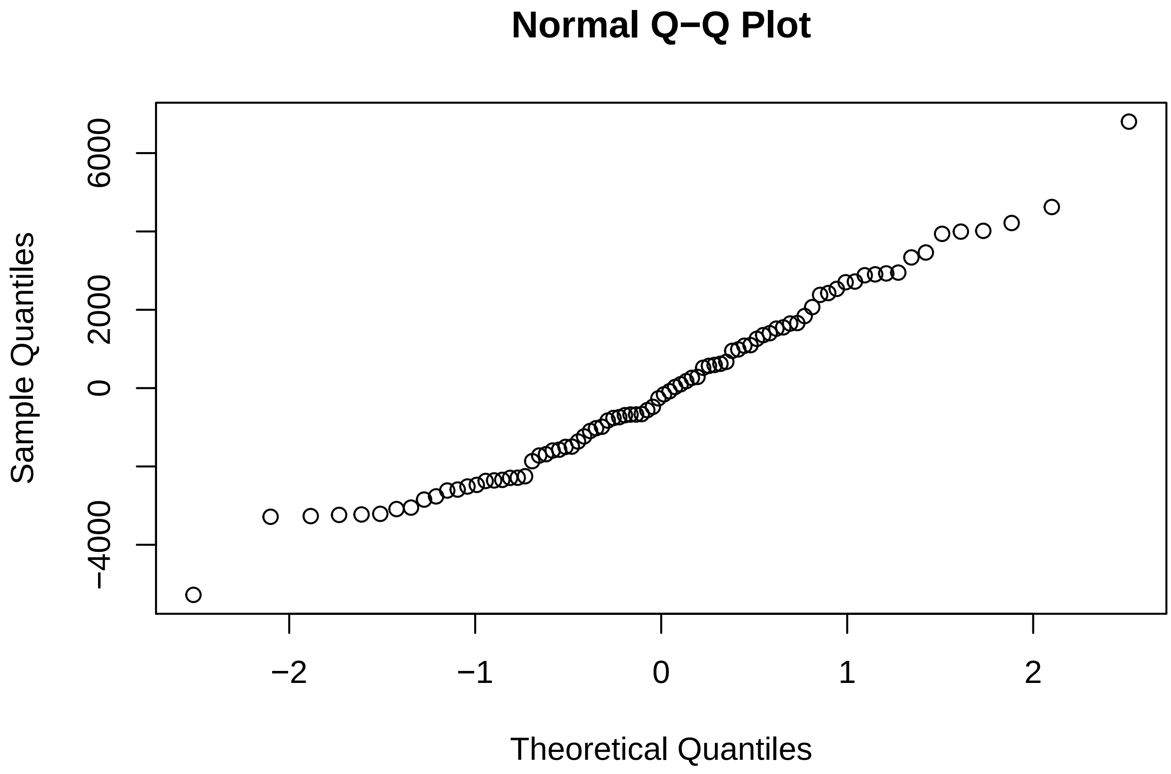
```
#Density does not resemble a normal curve.
```

```
#(c)
e <- resid(fit)
yhat <- fitted(fit)
plot(e ~ yhat)
lines(lowess(e ~ yhat))
```



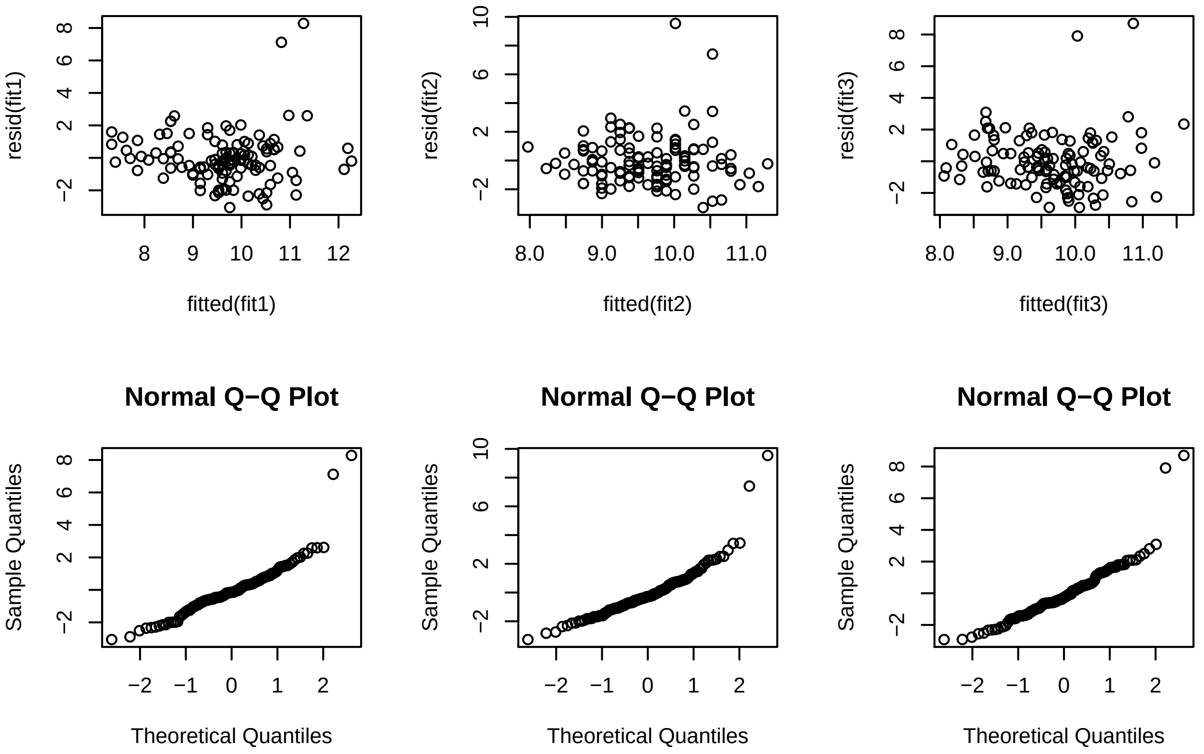
```
#Constant variance assumption seems fair.
```

```
#(d)
qqnorm(e)
```



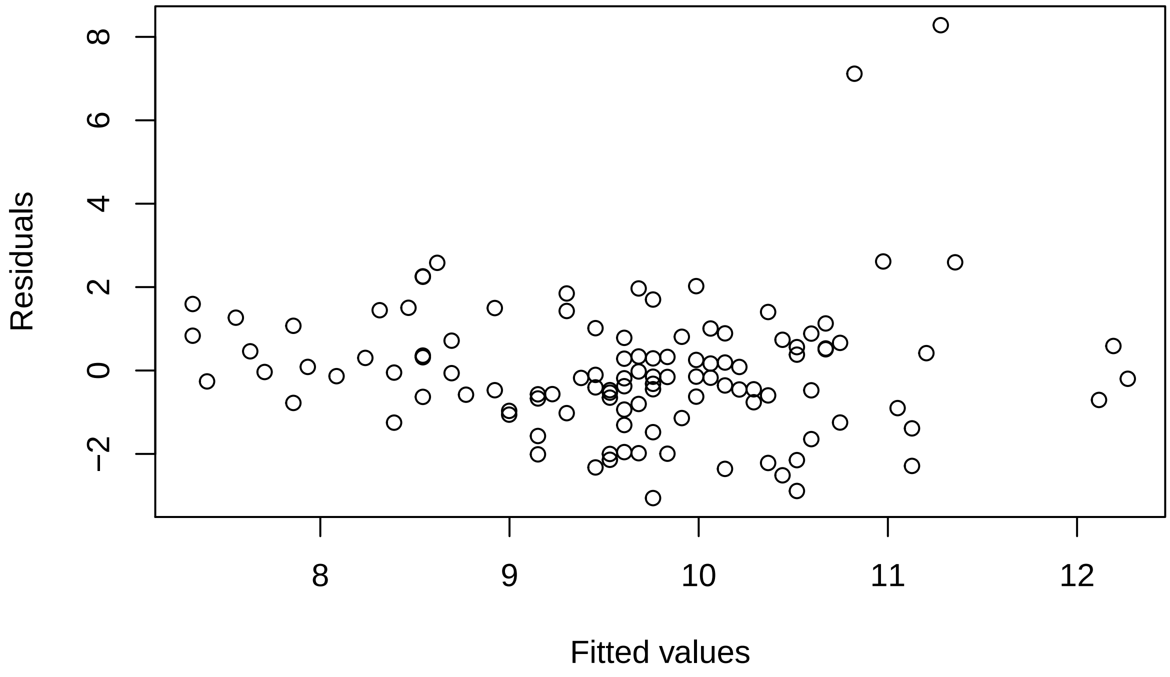
```
#It shows a normal probability plot.Normality assumption seems good.
```

```
#3.
#(a)
filename <- "~/Downloads/SENIC.txt"
SENIC <- read.table(file=filename, header=T)
fit1 <- lm(Stay ~ Risk, data=SENIC)
fit2 <- lm(Stay ~ AFS, data=SENIC)
fit3 <- lm(Stay ~ Xray, data=SENIC)
par(mfrow=c(2,3))
plot(resid(fit1) ~ fitted(fit1))
plot(resid(fit2) ~ fitted(fit2))
plot(resid(fit3) ~ fitted(fit3))
qqnorm(resid(fit1))
qqnorm(resid(fit2))
qqnorm(resid(fit3))
```



```
#These three plots show that the distribution of residuals are quite similar, and
#all of them have two outliers.
```

```
#(b)
x1<-fitted(fit1)
e<-resid(fit1)
plot(x1, e,xlab="Fitted values", ylab="Residuals")
identify(x1, e, n=2)
```



```
## integer(0)
```

```
#two outliers are 47 and 112.
summary(fit1)
```

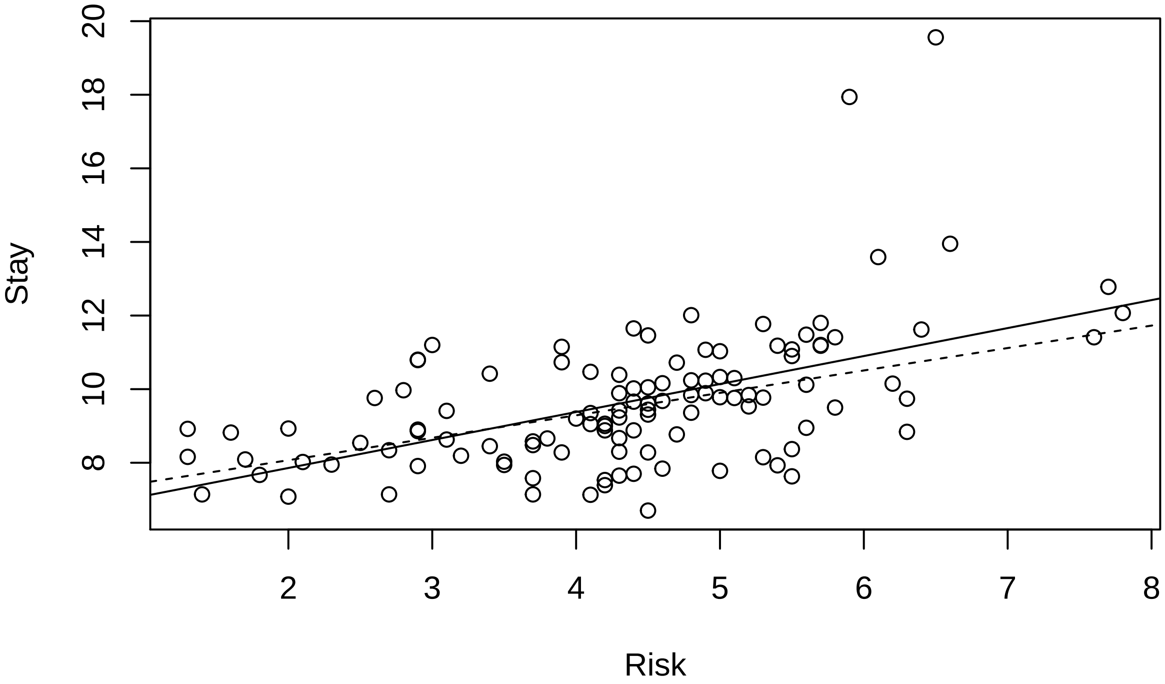
```
##
## Call:
## lm(formula = Stay ~ Risk, data = SENIC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0587 -0.7776 -0.1487  0.7159  8.2805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.3368     0.5213  12.156  < 2e-16 ***
## Risk          0.7604     0.1144   6.645 1.18e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 111 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2781
## F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09
```

```
refit <- update(fit1, subset=-c(47,112))
summary(refit)
```

```
##
## Call:
## lm(formula = Stay ~ Risk, data = SENIC, subset = -c(47, 112))
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89309 -0.67980 -0.08822  0.87180  3.07644
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.84922    0.40137  17.065  < 2e-16 ***
## Risk         0.60975    0.08881   6.866 4.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.238 on 109 degrees of freedom
## Multiple R-squared:  0.3019, Adjusted R-squared:  0.2955
## F-statistic: 47.14 on 1 and 109 DF,  p-value: 4.233e-10
```

```
plot(Stay ~ Risk, data=SENIC)
abline(fit1, lty=1)
abline(refit, lty=2)
```



```
#The distribution of residual is much normal after we remove that two outliers.
```

```
#(c)
SENIC[c(47,112),]
```

```
##      ID  Stay  Age Risk Cult  Xray Beds MS Reg Cen Nurses  AFS
## 47   47 19.56 59.9  6.5 17.2 113.7  306  2   1 273    172 51.4
## 112 112 17.94 56.2  5.9 26.4  91.8  835  1   1 791    407 62.9
```

```
predict(refit, SENIC[c(47,112),], interval="prediction")
```

```
##          fit      lwr      upr
## 47  10.81259 8.318631 13.30654
## 112 10.44674 7.966822 12.92665
```

```
#For hospital 47, the 95% PI is (8.3,13.3), the average length of stay in hospital 47 is 19.56,
#it is not in the prediction interval,hence it is a outlier.
#For hospital 112, the 95% PI is (7.9,12.9), the average length of stay in hospital 112 is 17.94,
#it is not in the prediction interval,hence it is a outlier.
```