Chapter 3 Solutions

LRPMFE

Question 1

(a)

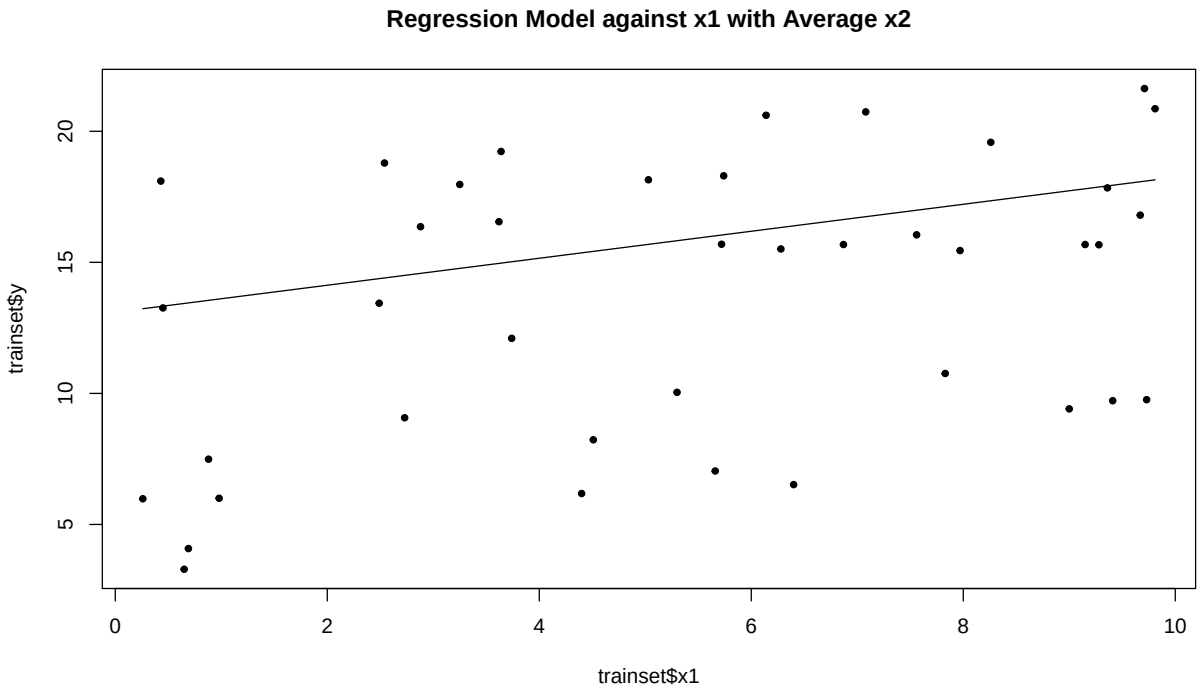
```
data=read.table("exercise2.1.DAT",header=T)
trainset=data.frame(data[1:40,])
testset=data.frame(data[41:60,2:3])
fit1a<-lm(y~x1+x2,trainset)
summary(fit1a)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = trainset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9585 -0.5865 -0.3356  0.3973  2.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31513    0.38769   3.392  0.00166 **
## x1           0.51481    0.04590  11.216 1.84e-13 ***
## x2           0.80692    0.02434  33.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 37 degrees of freedom
## Multiple R-squared:  0.9724, Adjusted R-squared:  0.9709
## F-statistic: 652.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

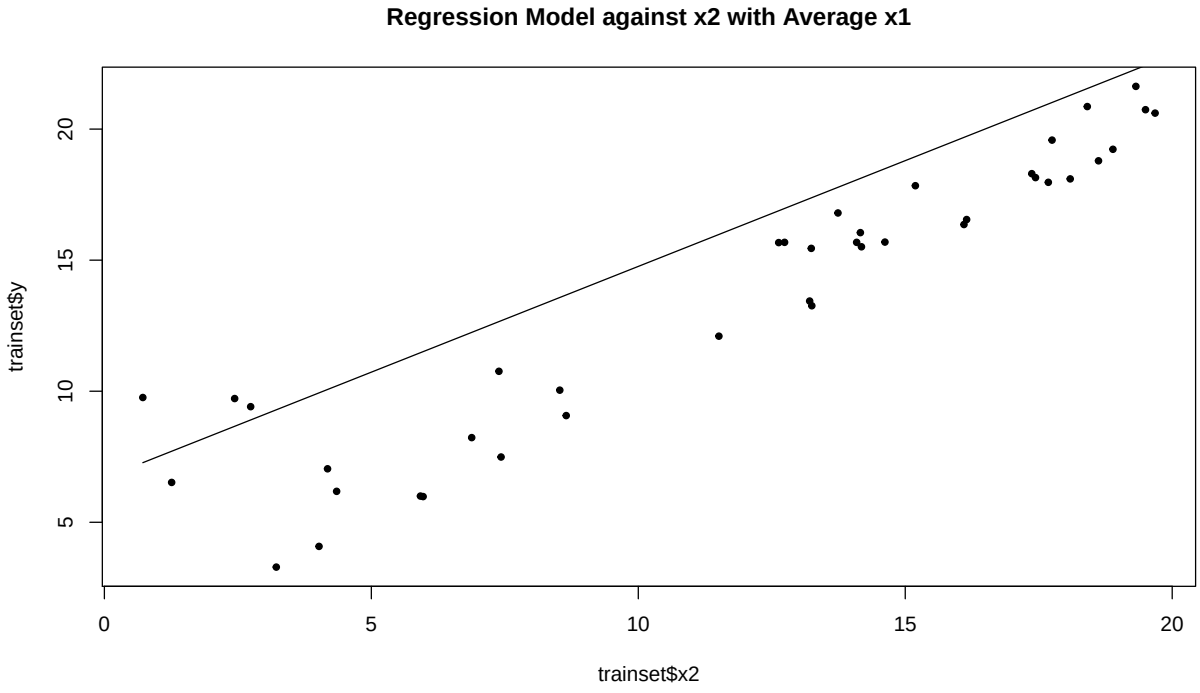
We can see that p-values for coefficients are small and coefficients are significant. RSE is 0.9 and R-squared is 0.974, implying a good fit.

(b)

```
plot(trainset$x1,trainset$y,main='Regression Model against x1 with Average x2',pch=20)
curve(fit1a$coefficients[1]+fit1a$coefficients[2]*x+mean(trainset$x2), add=TRUE)
```

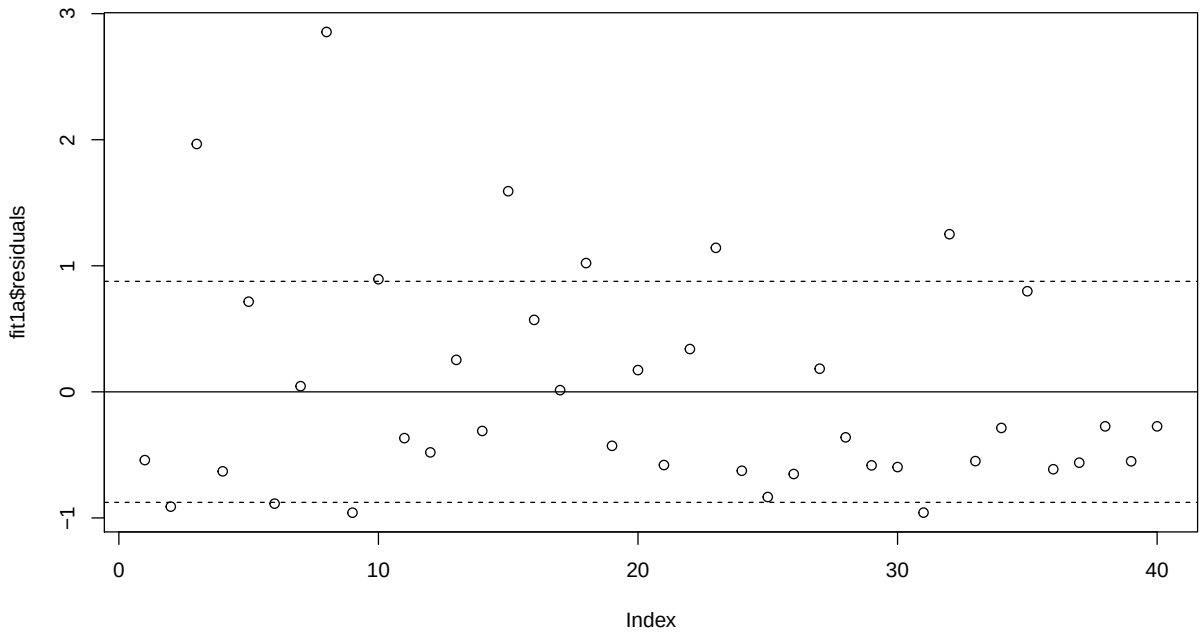



```
plot(trainset$x2,trainset$y,main='Regression Model against x2 with Average x1',pch=20)
curve(fit1a$coefficients[1]+fit1a$coefficients[3]*x+mean(trainset$x1), add=TRUE)
```



(c)

```r
resd<-sd(fit1a$residuals)
plot(fit1a$residuals)
abline(h=0)
abline(h=c(resd,-resd),lty=2)
```



We can see most of the residuals are within ± 1 standard-deviation bounds, no striking patterns.

(d)

```r
predict(fit1a,testset,interval="prediction", level=0.95)
```

```
##          fit       lwr       upr
## 41 14.812484 12.916966 16.708002
## 42 19.142865 17.241520 21.044211
## 43  5.916816  3.958626  7.875005
## 44 10.530475  8.636141 12.424809
## 45 19.012485 17.118597 20.906373
## 46 13.398863 11.551815 15.245911
## 47  4.829144  2.918323  6.739965
## 48  9.145767  7.228364 11.063170
## 49  5.892489  3.979060  7.805918
## 50 12.338639 10.426349 14.250929
## 51 18.908561 17.021818 20.795303
## 52 16.064649 14.212209 17.917088
## 53  8.963122  7.084081 10.842163
## 54 14.972786 13.094194 16.851379
## 55  5.859744  3.959679  7.759808
## 56  7.374900  5.480921  9.268879
```

```
## 57  4.535267  2.616996  6.453539
## 58 15.133280 13.282467 16.984094
## 59  9.100899  7.223395 10.978403
## 60 16.084900 14.196990 17.972810
```

Since testset y values are missing, we are not quite sure about the accuracy of predictions.

Question 2

(a)

The regression equation has the form $\ln \hat{y} = a + b \ln x$, which can be converted to $\hat{y} = e^a x^b$. The first order differential of y on x has x inside. So 'every increase of 1% in height corresponds to a predicted increase of 0.8% in earnings' is a bit confusing here because the slope depends on x. We assume the slope is regarding to the log values, so b=0.8. From $30000 = e^a 66^{0.8}$ we can solve $a = 6.96$. Thus$\ln \hat{y} = 6.96 + 0.8 \ln x$.

The 95% confidence interval for predictions should be $\ln \bar{y} \pm 2\sigma$. Since $\bar{y}/1.1 < y < 1.1\bar{y}$, we have $\ln \bar{y} - \ln 1.1 < y < \ln \bar{y} + \ln 1.1$. So $\ln 1.1 = 2\sigma$ and we get $\sigma = 0.047655$

(b)

$SSR = \sum(\ln \hat{y} - \ln \bar{y})^2 = \sum(0.8 \ln x - 0.8 \ln \bar{x})^2 = 0.64 * 0.05^2 n = 0.0016n$ $R^2 = \frac{SSR}{SST} = \frac{0.0016}{0.0016 + 0.047655^2} = 0.4133$

Question 3

```
var1<-rnorm(1000,0,1)
var2<-rnorm(1000,0,1)
fit3a<-lm(var1~var2)
summary(fit3a)
```

```
##
## Call:
## lm(formula = var1 ~ var2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2565 -0.6703 -0.0315  0.6727  3.4367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02874    0.03283   0.875    0.382
## var2         0.01430    0.03209   0.446    0.656
##
## Residual standard error: 1.037 on 998 degrees of freedom
## Multiple R-squared:  0.0001991,  Adjusted R-squared:  -0.0008027
## F-statistic: 0.1987 on 1 and 998 DF,  p-value: 0.6558
```

Pvalue for slope is large so it's not significant.

(b)

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/summary(fit)$coefficients[2,2]
}
sum(z.scores>2)
```

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

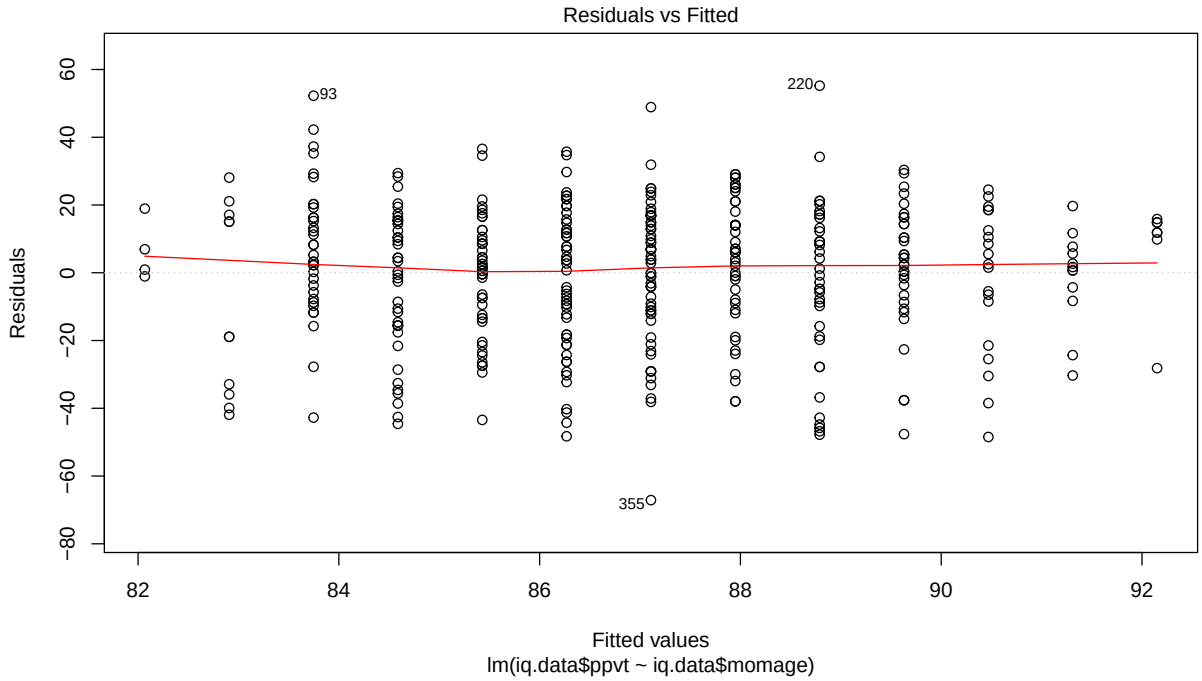
Question 4

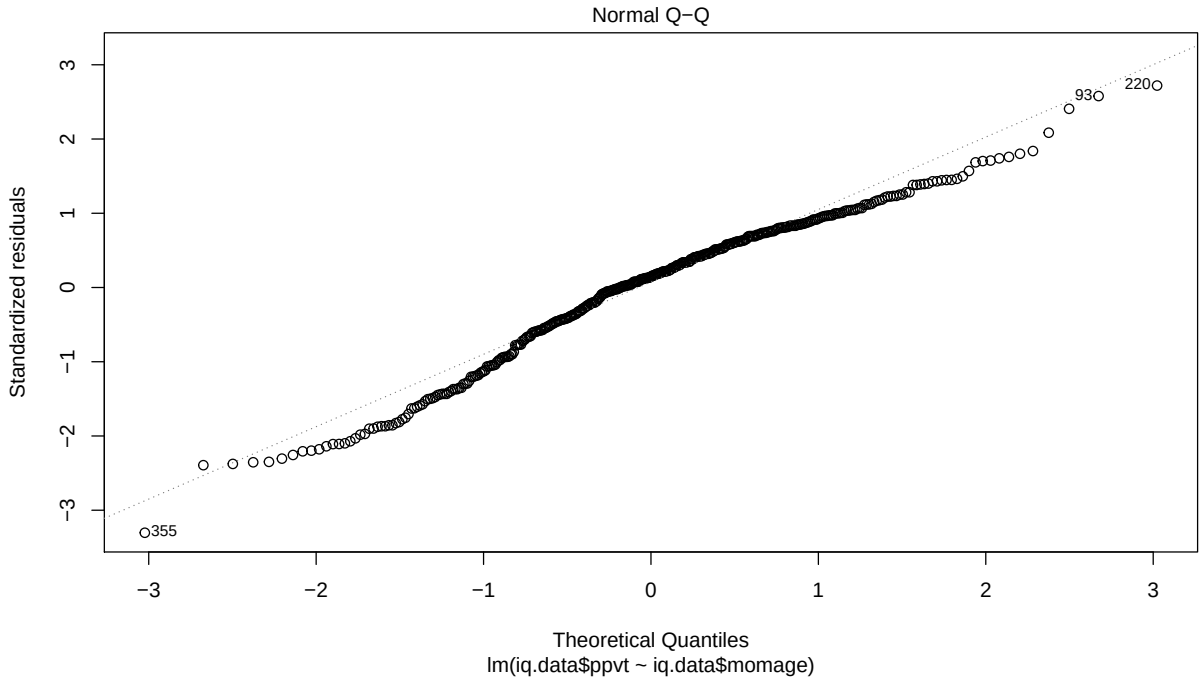
(a)

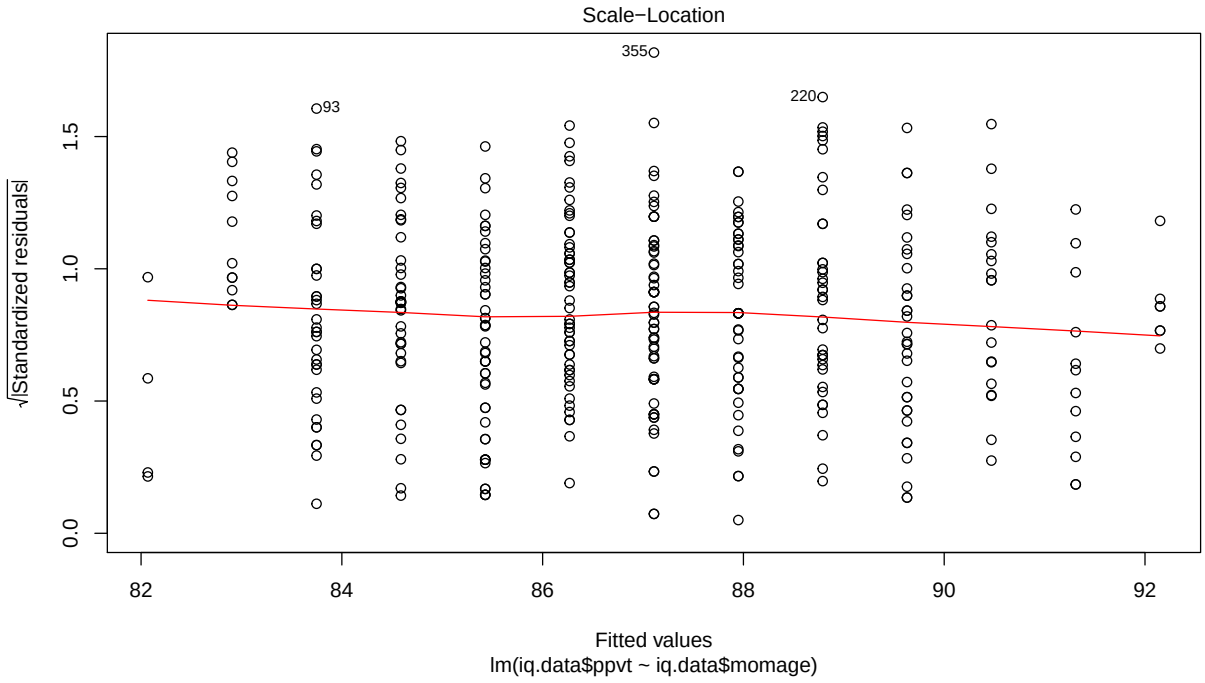
```
library("foreign")
iq.data<-read.dta("child.iq.dta")
fit4a<-lm(iq.data$ppvt~iq.data$momage)
summary(fit4a)
```

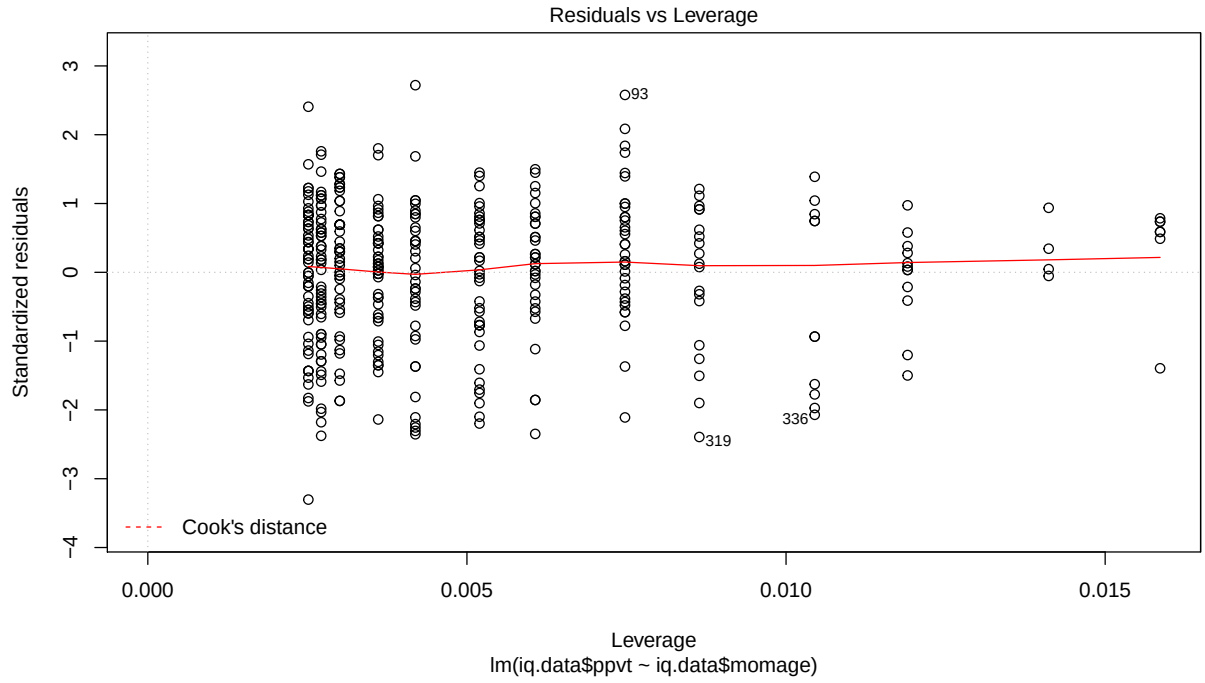
```
##
## Call:
## lm(formula = iq.data$ppvt ~ iq.data$momage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.109 -11.798   2.971  14.860  55.210
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     67.7827     8.6880   7.802 5.42e-14 ***
## iq.data$momage   0.8403     0.3786   2.219    0.027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.34 on 398 degrees of freedom
## Multiple R-squared:  0.01223,    Adjusted R-squared:  0.009743
## F-statistic: 4.926 on 1 and 398 DF,  p-value: 0.02702
```

```
plot(fit4a)
```

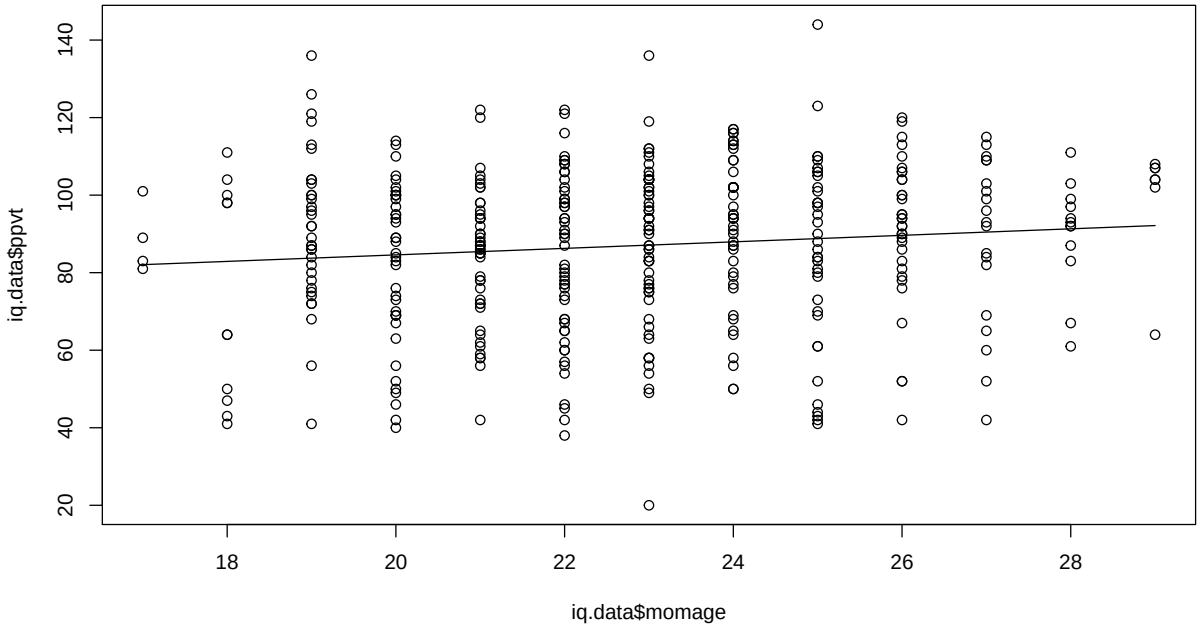

Residuals vs Fitted
Residuals
Fitted values
lm(iq.data$ppvt ~ iq.data$momage)
93
220
355


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(iq.data$ppvt ~ iq.data$momage)
93
220
355






```
plot(iq.data$momage,iq.data$ppvt)
curve(fit4a$coefficients[1]+fit4a$coefficients[2]*x,add=TRUE)
```



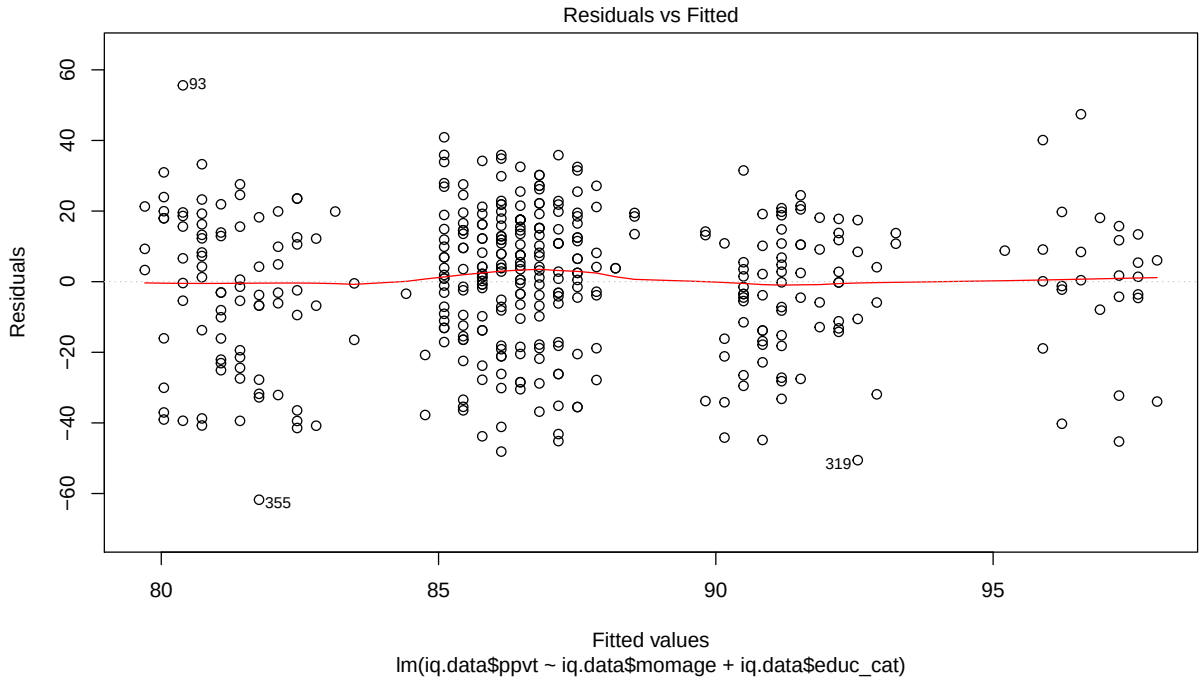
The slope coefficient is 0.84 and p-value is 0.027. The residuals graph shows that they don't have heteroskedasticity and can be regarded as normally distributed. Since the slope is positive, if assuming no other factors will influence children's scores, the older to give birth, the higher for the test score.

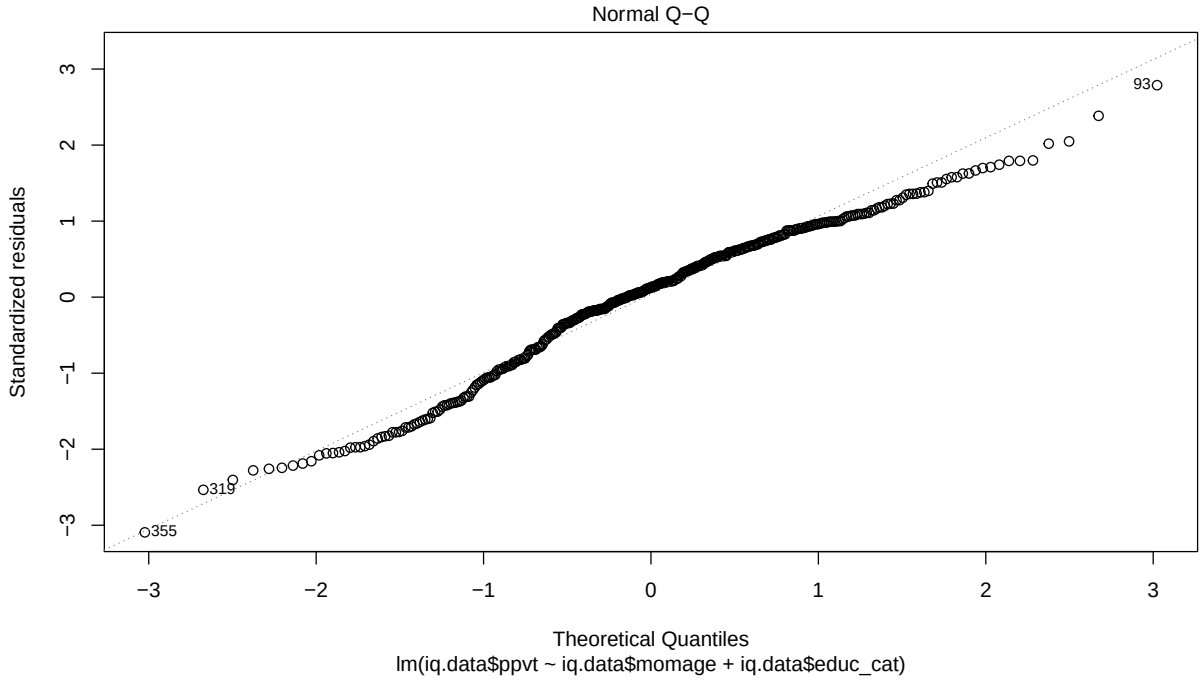
(b)

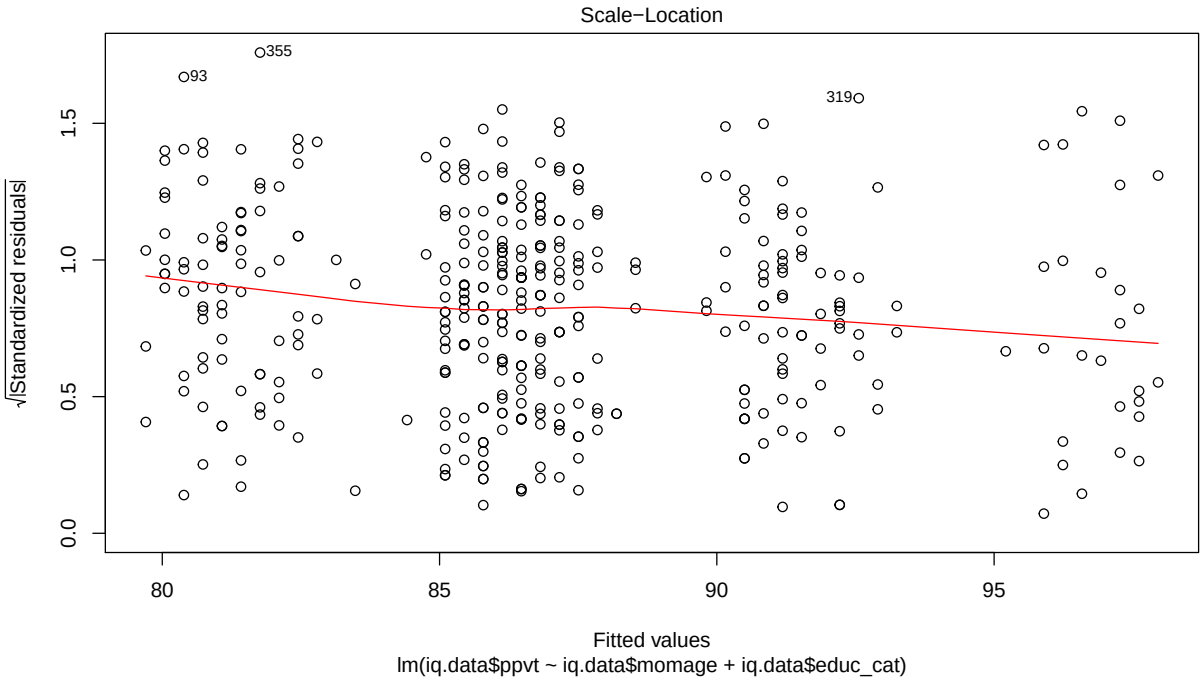
```
fit4b<-lm(iq.data$ppvt~iq.data$momage+iq.data$educ_cat)
summary(fit4b)
```

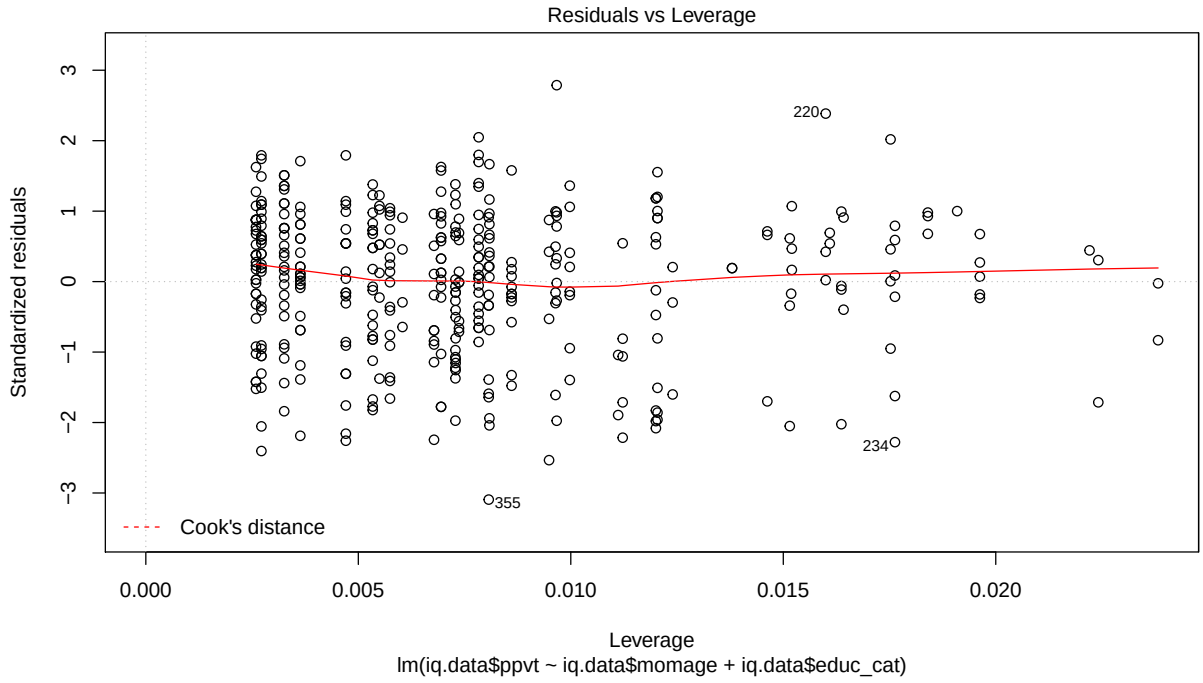
```
##
## Call:
## lm(formula = iq.data$ppvt ~ iq.data$momage + iq.data$educ_cat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.763 -13.130   2.495  14.620  55.610
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       69.1554     8.5706   8.069 8.51e-15 ***
## iq.data$momage     0.3433     0.3981   0.862 0.389003
## iq.data$educ_cat   4.7114     1.3165   3.579 0.000388 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.05 on 397 degrees of freedom
## Multiple R-squared:  0.04309,    Adjusted R-squared:  0.03827
## F-statistic: 8.939 on 2 and 397 DF,  p-value: 0.0001594
```

```
plot(fit4b)
```


Residuals vs Fitted
Residuals
Fitted values
lm(iq.data$ppvt ~ iq.data$momage + iq.data$educ_cat)
93
355
319
80
85
90
95
60
40
20
0
−20
−40
−60

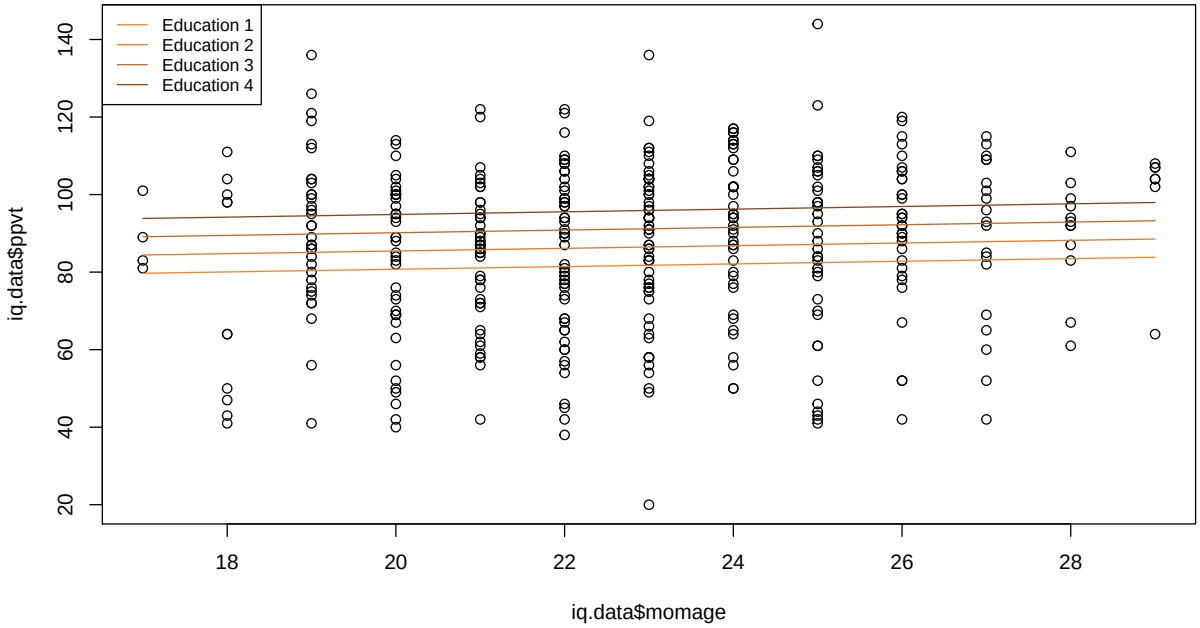

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(iq.data$ppvt ~ iq.data$momage + iq.data$educ_cat)
93
319
355
−3
−2
−1
0
1
2
3






```
color4b=c('chocolate1','chocolate2','chocolate3','chocolate4')
plot(iq.data$momage,iq.data$ppvt)
for(i in 1:4){
  curve(cbind(1,x,i) %*% fit4b$coefficients, col=color4b[i] ,add=TRUE)
}
legend("topleft", legend=c("Education 1", "Education 2","Education 3","Education 4"),
```

```
       col=color4b, lty=1, cex=0.8)
```



The coefficients for momage and education are 0.34 and 4.7. P-value for momage is 0.39, so momage is not significant. We also find that higher education tends to have higher test scores. For the best timing of birth, it is the same as in (a)

(c)

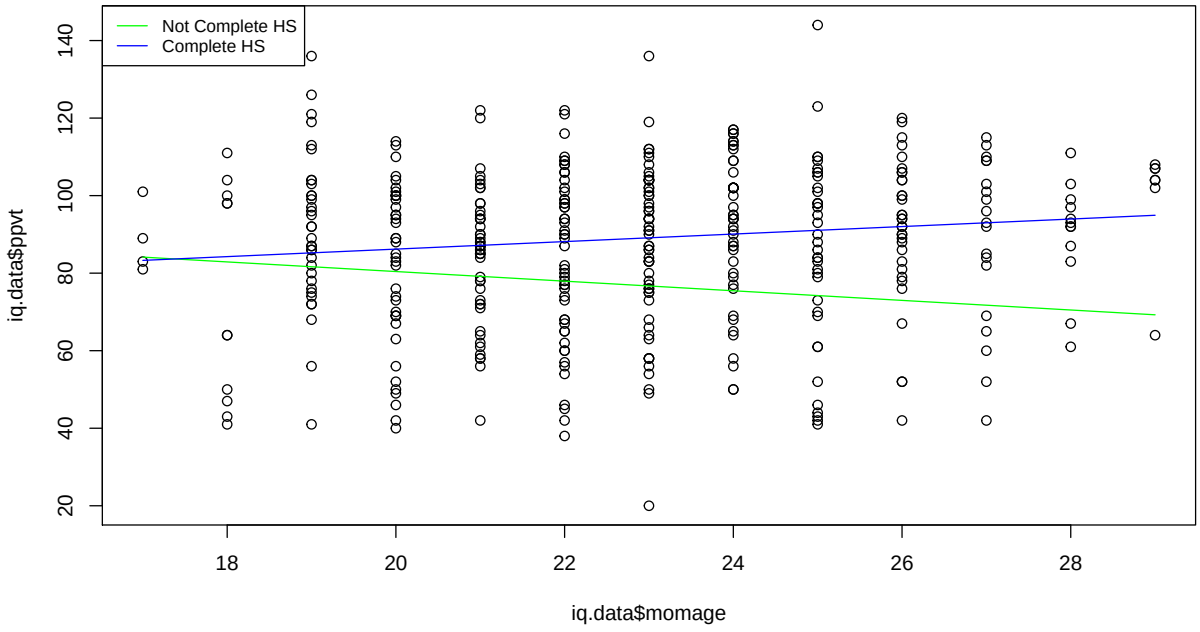
Consider those with education indices >=2 complete the high school.

```
iq.data$hs<-ifelse(iq.data$educ_cat>=2,1,0)
fit4c<-lm(ppvt~momage+hs+hs*momage,iq.data)
summary(fit4c)
```

```
##
## Call:
## lm(formula = ppvt ~ momage + hs + hs * momage, data = iq.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.696 -12.407   2.022  14.804  54.343
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 105.2202    17.6454   5.963 5.49e-09 ***
## momage       -1.2402     0.8113  -1.529   0.1271
## hs          -38.4088    20.2815  -1.894   0.0590 .
## momage:hs     2.2097     0.9181   2.407   0.0165 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.85 on 396 degrees of freedom
## Multiple R-squared:  0.06417,    Adjusted R-squared:  0.05708
## F-statistic: 9.051 on 3 and 396 DF,  p-value: 8.276e-06
```

```
plot(iq.data$momage,iq.data$ppvt)
curve(cbind(1,x,0,0) %*% fit4c$coefficients, col='green' ,add=TRUE)
curve(cbind(1,x,1,x) %*% fit4c$coefficients, col='blue' ,add=TRUE)
legend("topleft", legend=c("Not Complete HS", "Complete HS"),
       col=c("green", "blue"), lty=1, cex=0.8)
```



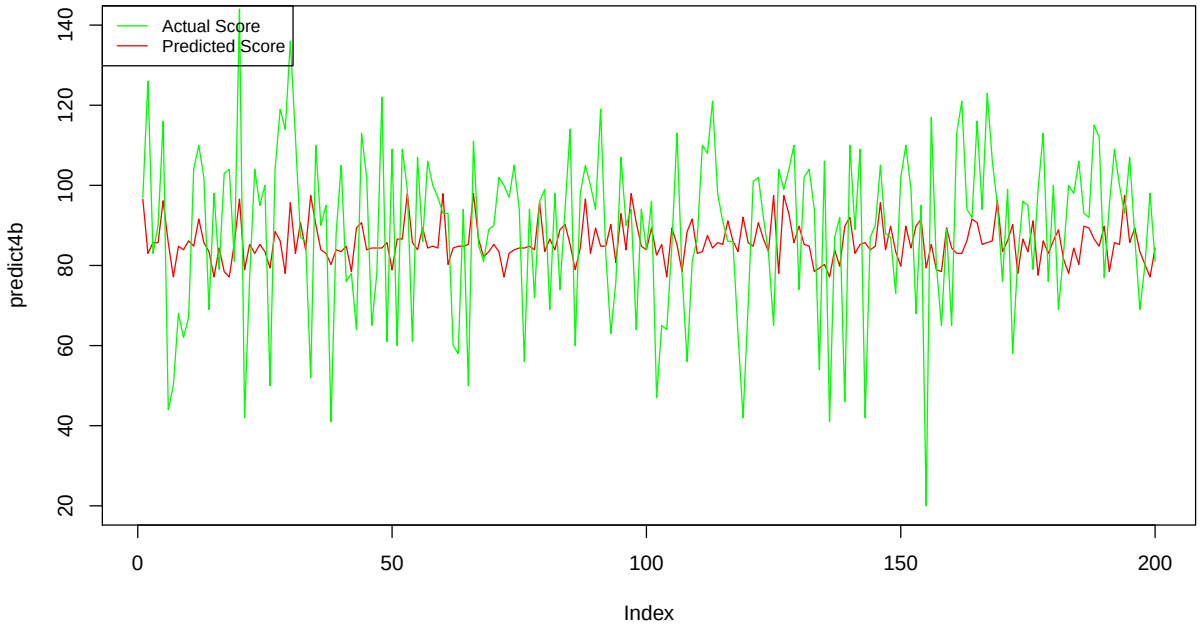
(d)

```
fit4d<-lm(ppvt~momage+educ_cat,iq.data[1:200,])
summary(fit4d)
```

```
##
## Call:
## lm(formula = ppvt ~ momage + educ_cat, data = iq.data[1:200,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.358 -12.967   2.866  14.435  58.428
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.6295    11.8202   5.383 2.07e-07 ***
## momage        0.4473     0.5516   0.811  0.41836
## educ_cat      5.4434     1.8228   2.986  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.58 on 197 degrees of freedom
## Multiple R-squared:  0.06199,    Adjusted R-squared:  0.05246
## F-statistic: 6.509 on 2 and 197 DF,  p-value: 0.001831
```

```
testdata=iq.data[201:400,c('momage','educ_cat')]
predict4b=predict(fit4d,testdata)
plot(predict4b,type='l',col='red',ylim =c(20,140))
lines(iq.data[201:400,]$ppvt,col='green',ylim=c(20,140))
legend("topleft", legend=c("Actual Score", "Predicted Score"),
       col=c("green", "red"), lty=1, cex=0.8)
```



We can see the predicted results are not satisfactory.

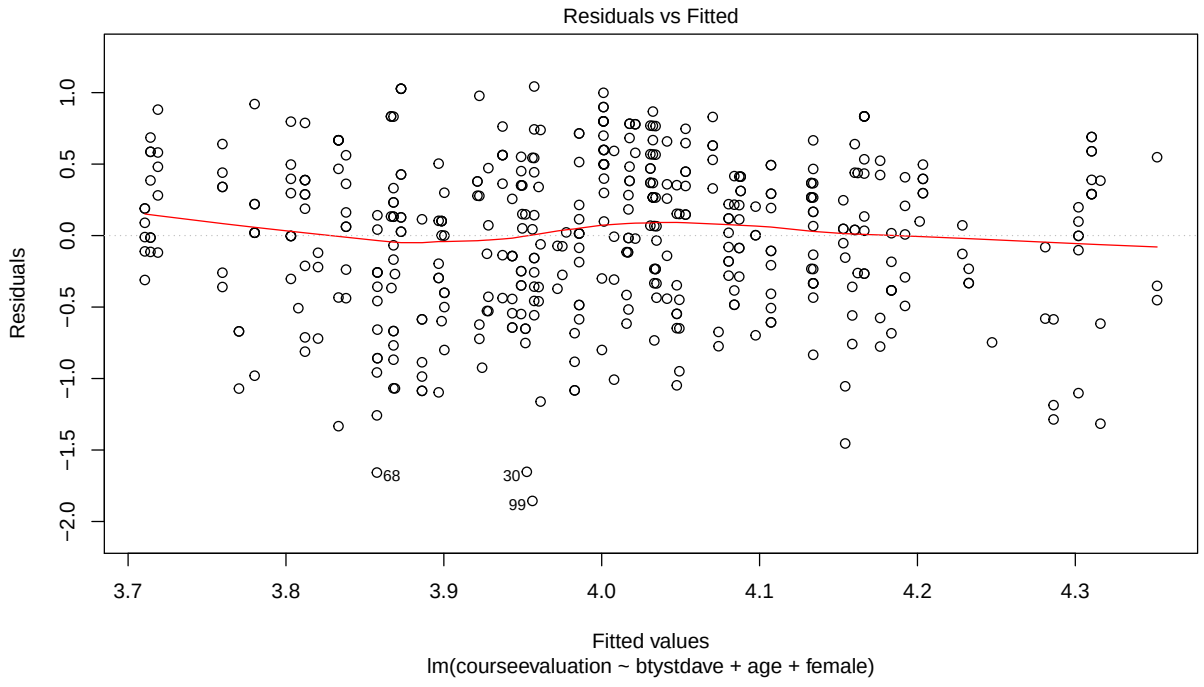
Question 5

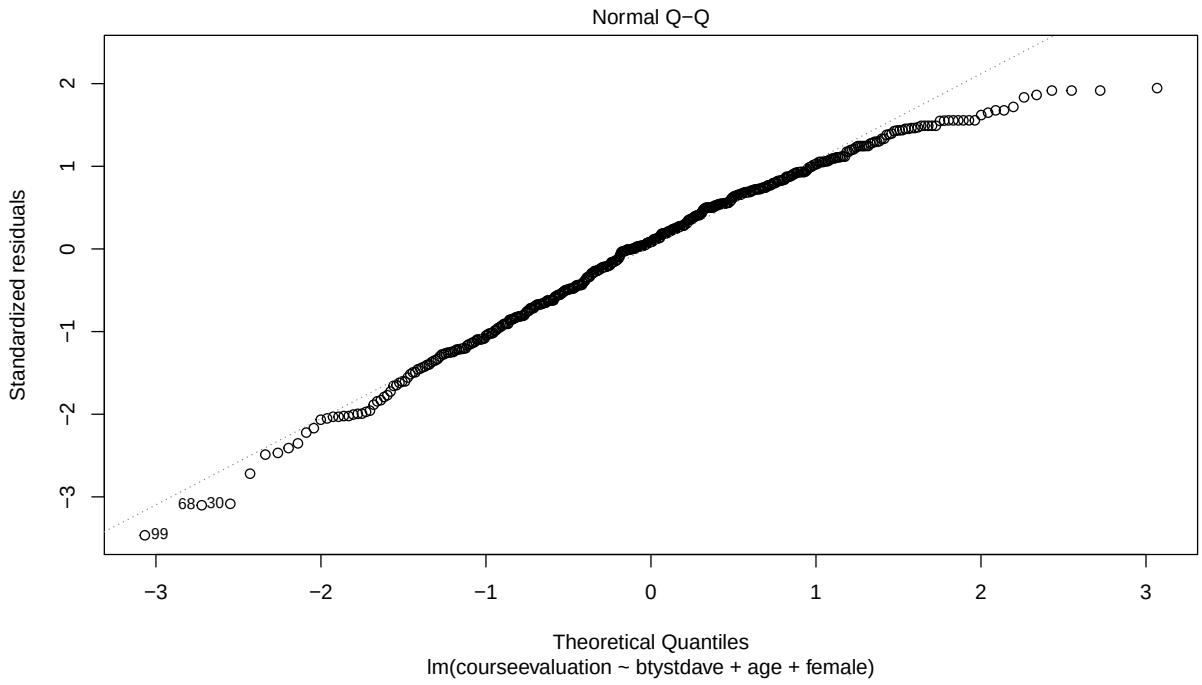
(a)

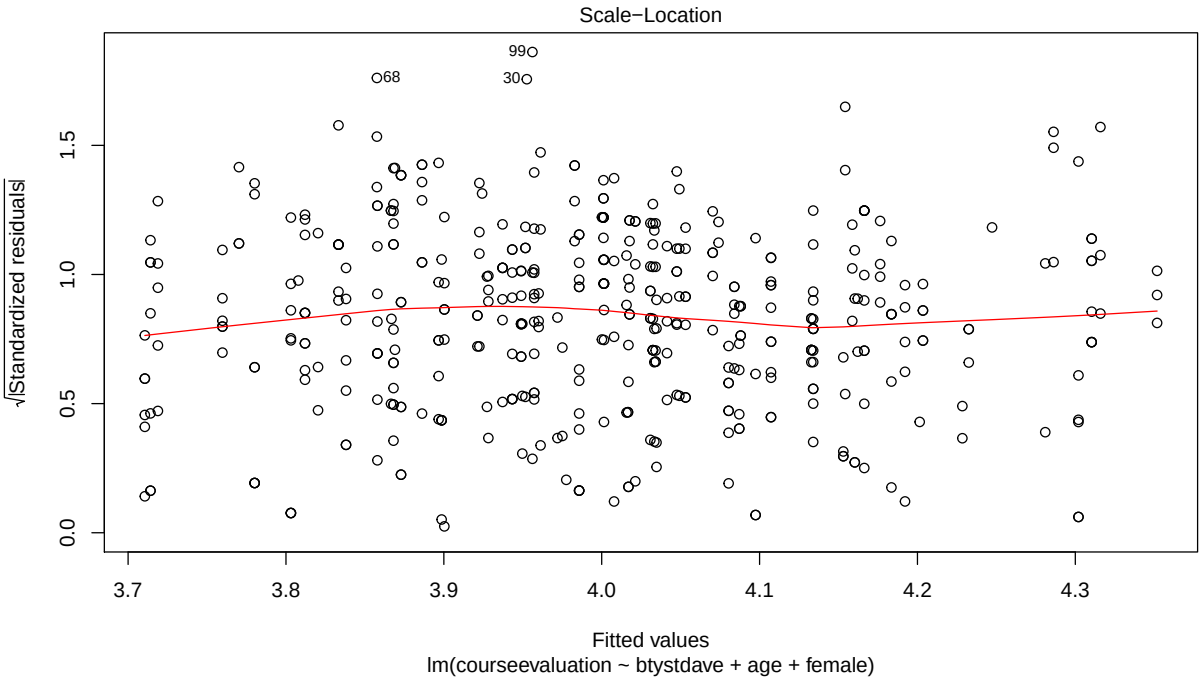
```
library(readxl)
data<-read_xls('ProfEvaltnsBeautyPublic.xls')
fit5a<-lm(courseevaluation~btystdave+age+female,data)
summary(fit5a)
```

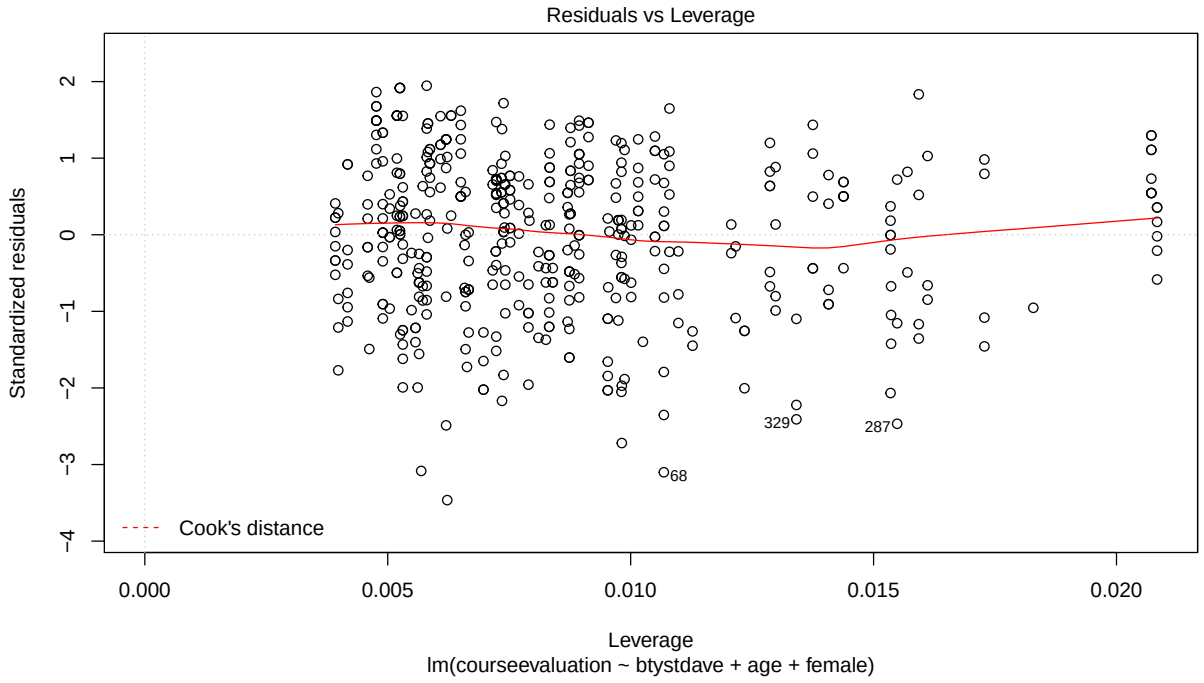
```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + age + female, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85612 -0.35831  0.04697  0.39308  1.04276
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.225242   0.142820  29.584  < 2e-16 ***
## btystdave    0.139978   0.033243   4.211 3.06e-05 ***
## age         -0.002602   0.002768  -0.940    0.348
## female      -0.210792   0.052824  -3.990 7.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5374 on 459 degrees of freedom
## Multiple R-squared:  0.06809,    Adjusted R-squared:  0.062
## F-statistic: 11.18 on 3 and 459 DF,  p-value: 4.305e-07
```

```
plot(fit5a)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(courseevaluation ~ btystdave + age + female)
99
68
30


Scale−Location
√|Standardized residuals|
Fitted values
lm(courseevaluation ~ btystdave + age + female)
99
68
30

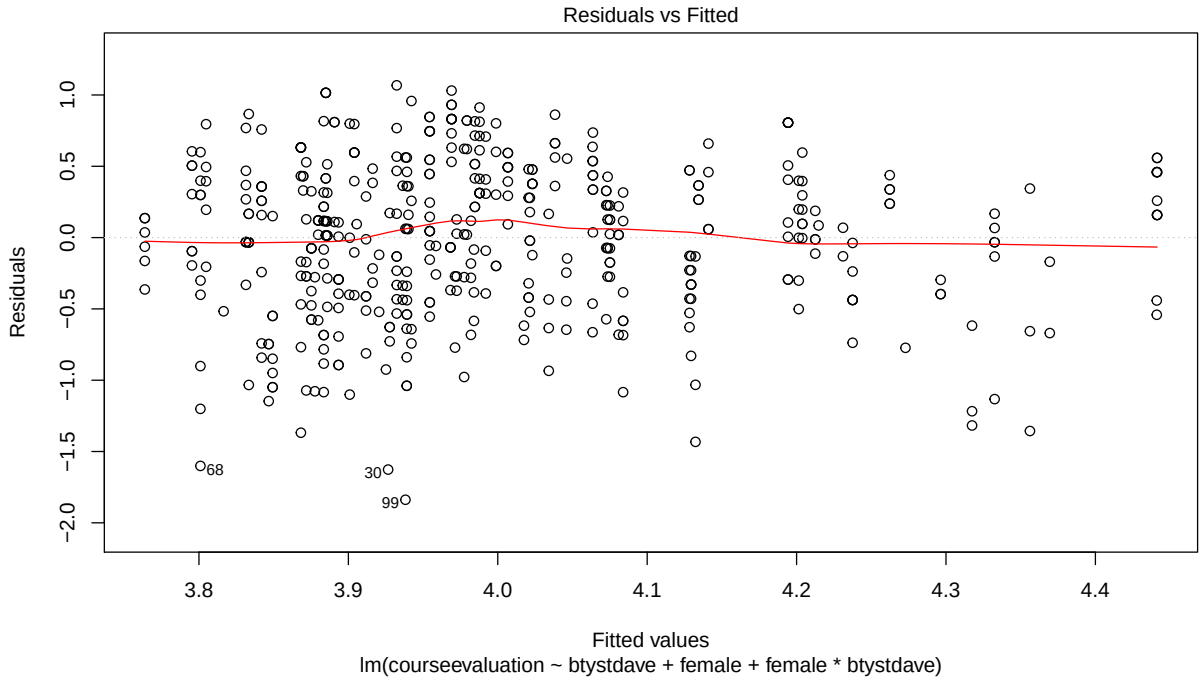


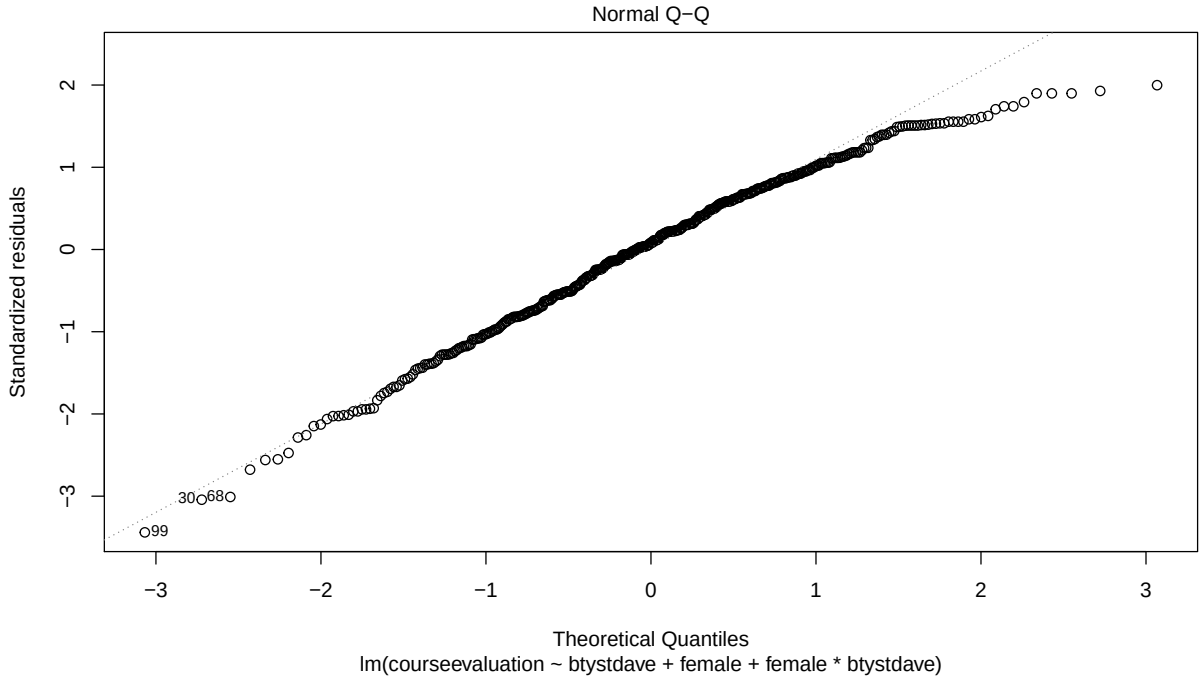
(b)

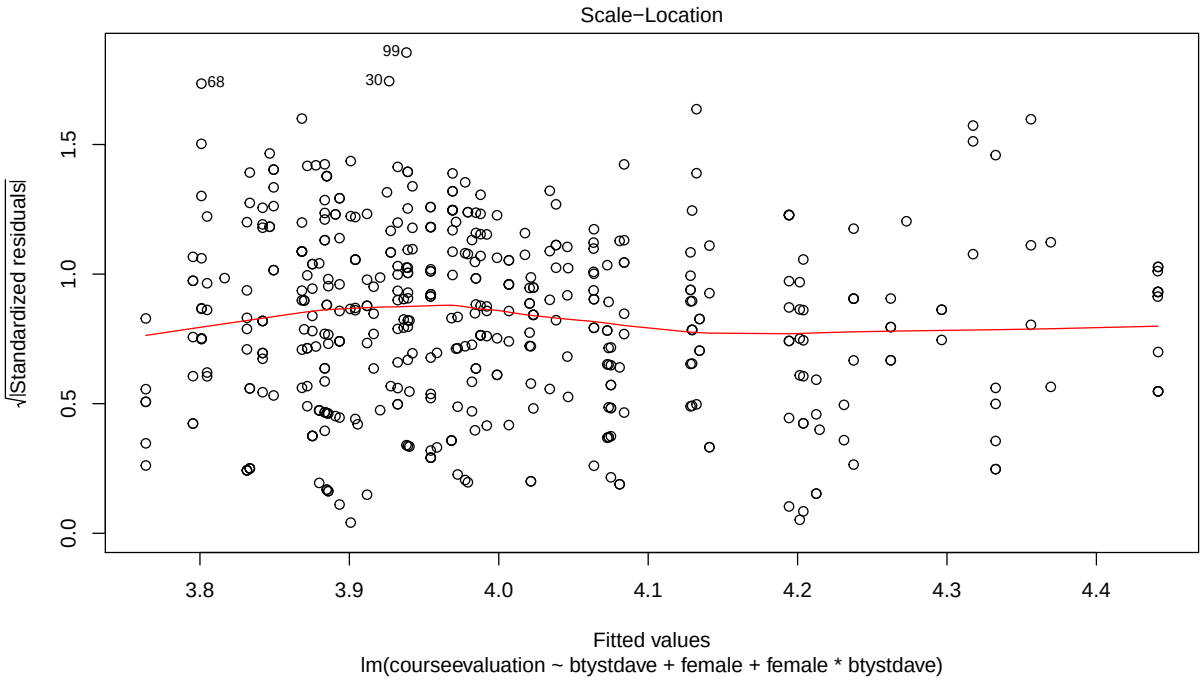
```
fit5b<-lm(courseevaluation~btystdave+female+female*btystdave,data)
summary(fit5b)
```

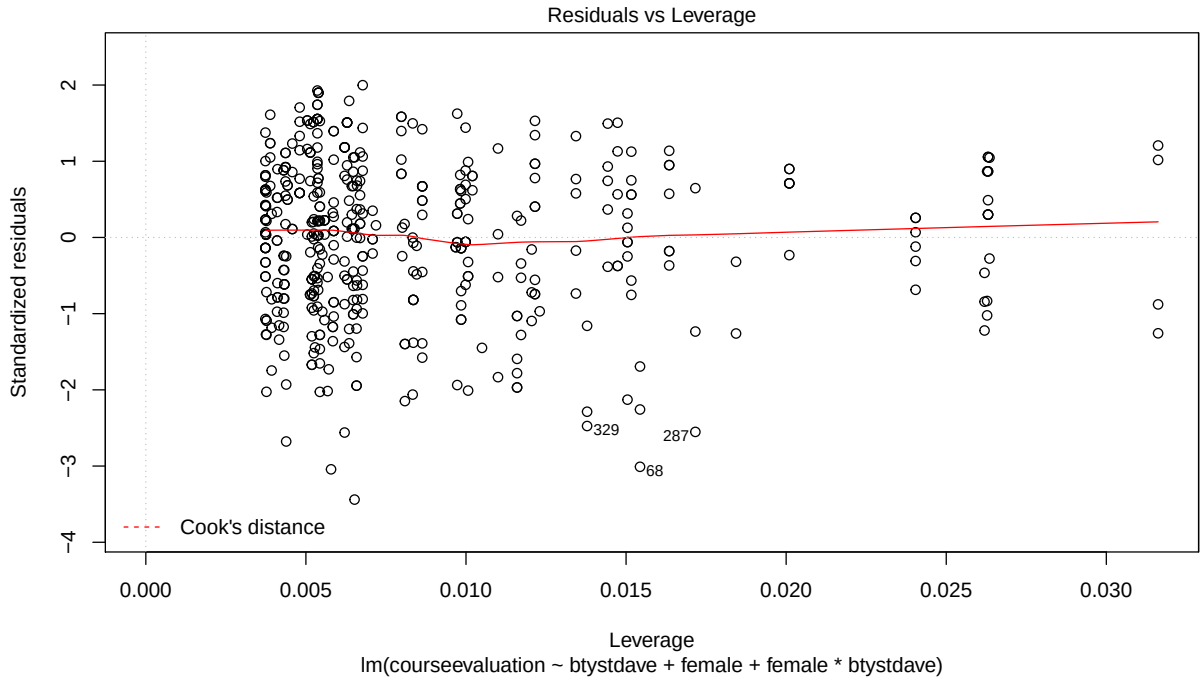
```
##
## Call:
## lm(formula = courseevaluation ~ btystdave + female + female *
##     btystdave, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83820 -0.37387  0.04551  0.39876  1.06764
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.10364    0.03359 122.158  < 2e-16 ***
## btystdave          0.20027    0.04333   4.622 4.95e-06 ***
## female            -0.20505    0.05103  -4.018 6.85e-05 ***
## btystdave:female  -0.11266    0.06398  -1.761   0.0789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5361 on 459 degrees of freedom
## Multiple R-squared:  0.07256,    Adjusted R-squared:  0.0665
## F-statistic: 11.97 on 3 and 459 DF,  p-value: 1.471e-07
```

```
plot(fit5b)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
68
30
99
3.8
3.9
4.0
4.1
4.2
4.3
4.4
Fitted values
lm(courseevaluation ~ btystdave + female + female * btystdave)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
30
68
99
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(courseevaluation ~ btystdave + female + female * btystdave)


Scale−Location
99
30
68
√|Standardized residuals|
0.0
0.5
1.0
1.5
3.8
3.9
4.0
4.1
4.2
4.3
4.4
Fitted values
lm(courseevaluation ~ btystdave + female + female * btystdave)


Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
−3
−4
329
287
68
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(courseevaluation ~ btystdave + female + female * btystdave)