ADA HW5

Yan Wu (Uni: yw2592)

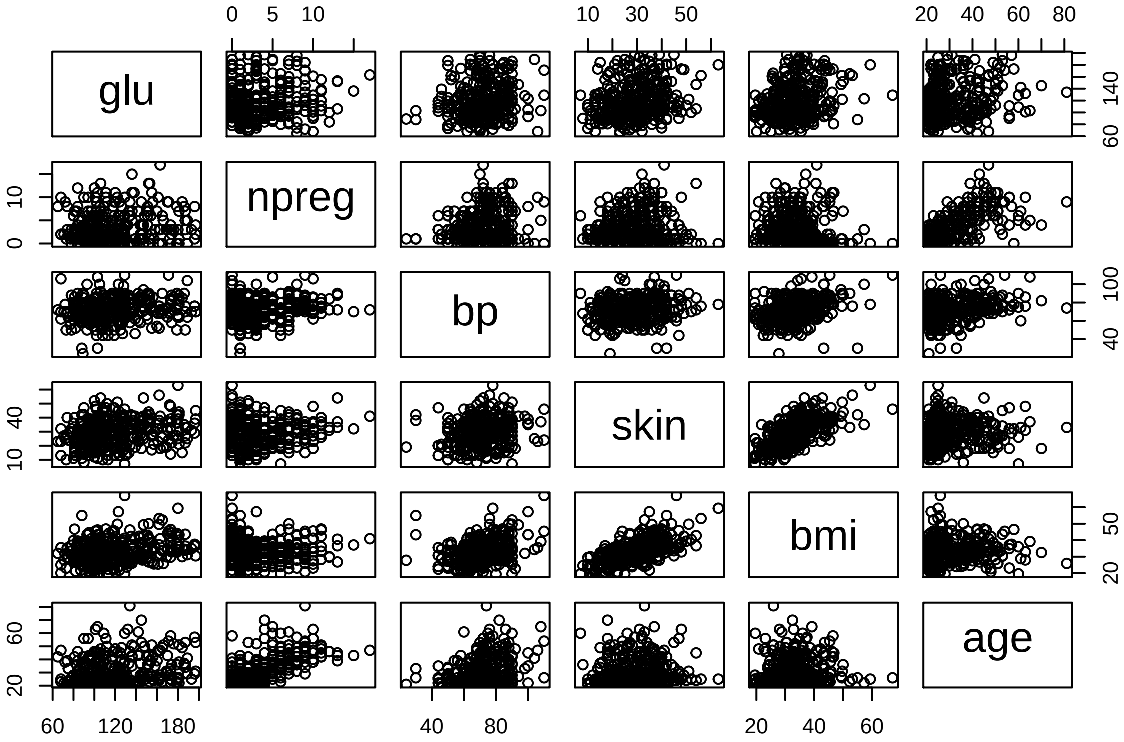
Friday, October 10, 2014

Problem 1

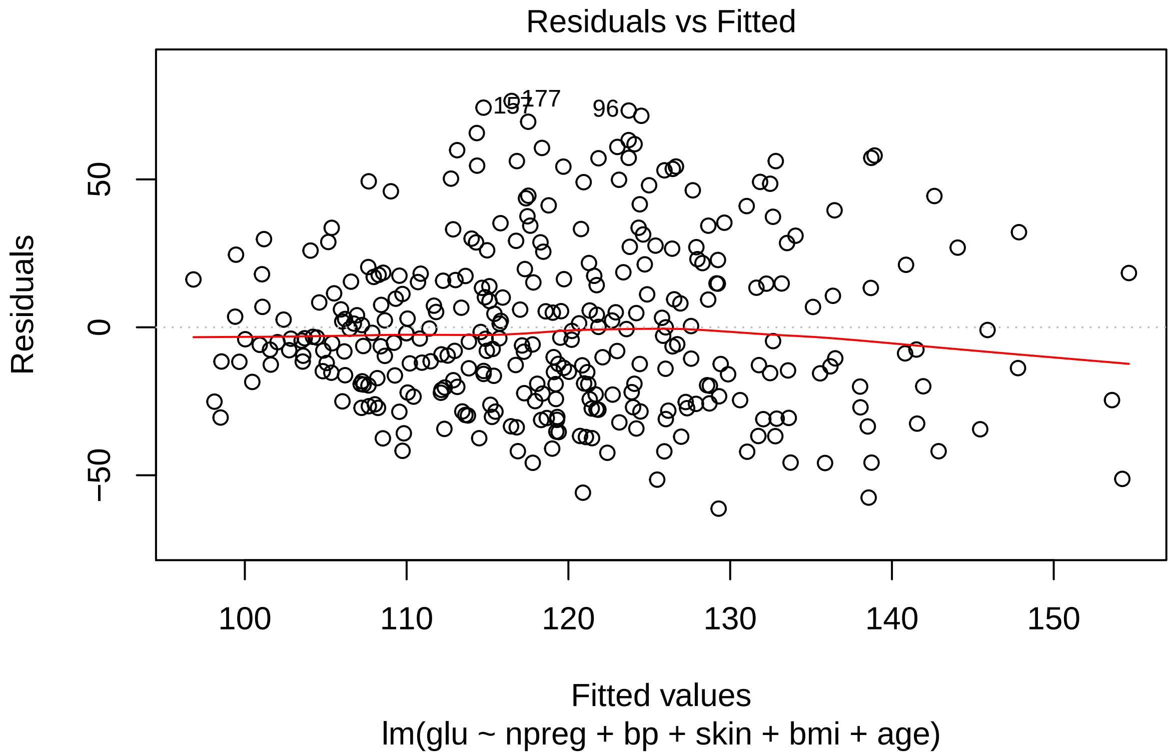
```
require(MASS)
```

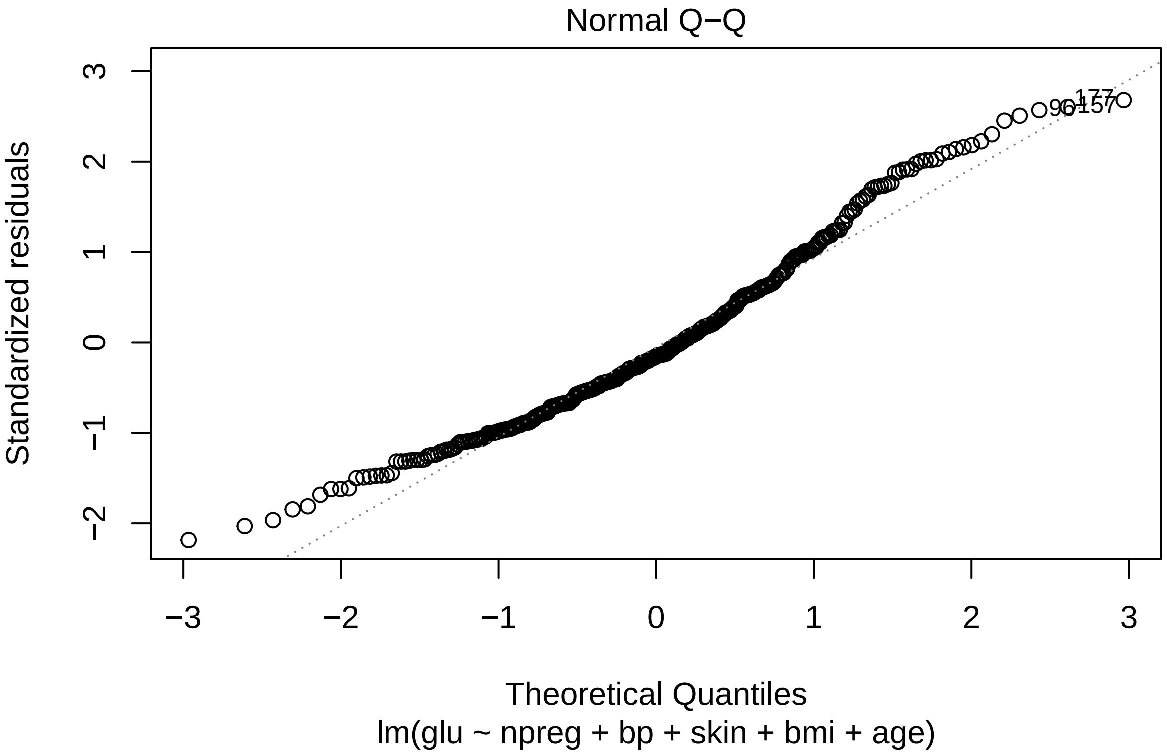
```
## Loading required package: MASS
```

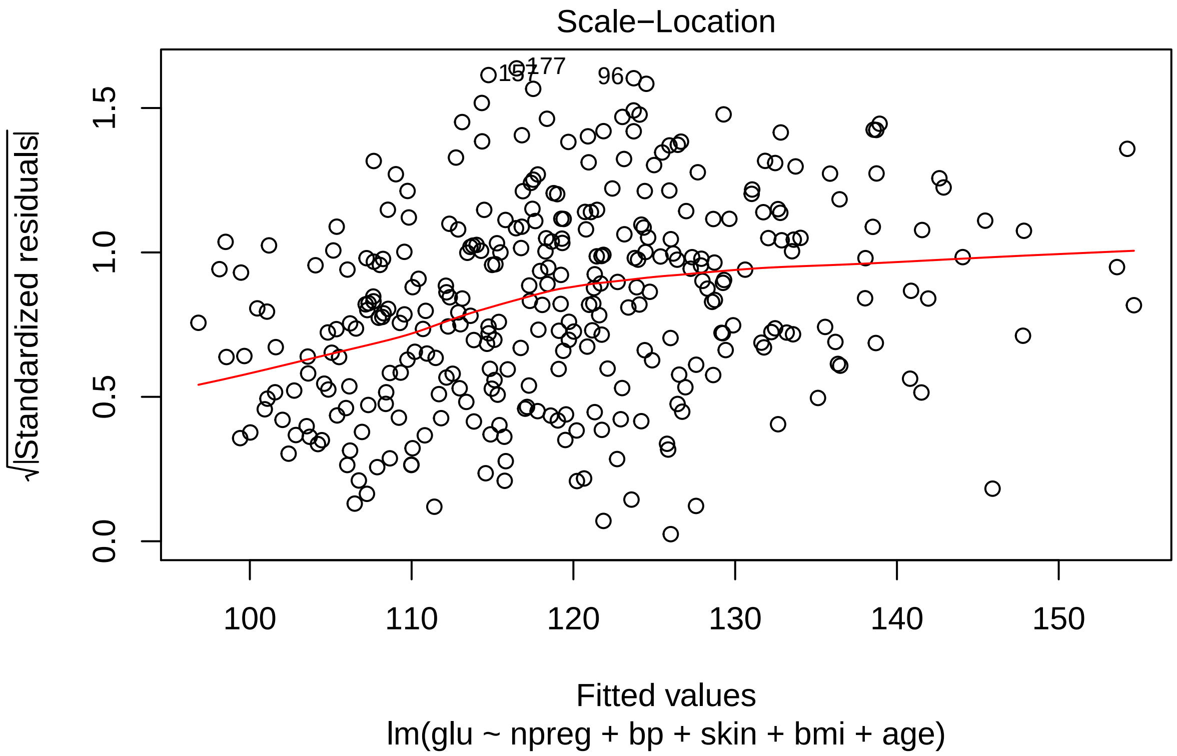
```
data <- Pima.te
pairs(~glu+npreg+bp+skin+bmi+age,data=data)
```

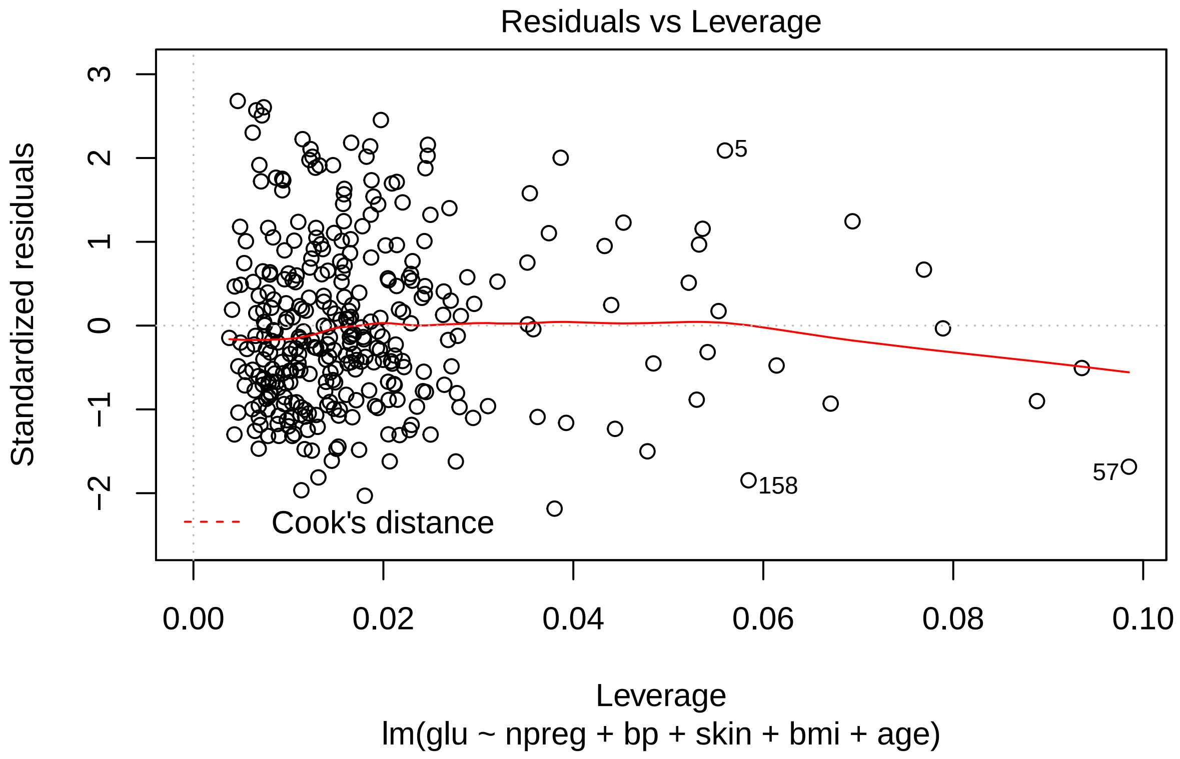


```
fit1 <- lm(glu~npreg+bp+skin+bmi+age,data=data)
plot(fit1)
```


Residuals vs Fitted
157
177
96
Residuals
50
0
−50
100
110
120
130
140
150
Fitted values
lm(glu ~ npreg + bp + skin + bmi + age)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
177
96
157
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(glu ~ npreg + bp + skin + bmi + age)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
100
110
120
130
140
150
Fitted values
lm(glu ~ npreg + bp + skin + bmi + age)
157
177
96



```r
summary(fit1)
```

```
##
## Call:
## lm(formula = glu ~ npreg + bp + skin + bmi + age, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61.29 -20.56  -4.36  17.37  76.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.831     10.309    5.51  7.2e-08 ***
## npreg         -0.875      0.647   -1.35  0.17735
## bp             0.104      0.138    0.75  0.45353
## skin           0.263      0.216    1.21  0.22575
## bmi            0.796      0.302    2.64  0.00880 **
## age            0.764      0.207    3.69  0.00026 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.6 on 326 degrees of freedom
## Multiple R-squared:  0.134,  Adjusted R-squared:  0.121
## F-statistic: 10.1 on 5 and 326 DF,  p-value: 5.58e-09
```

From the result of the fitted model, we have the average estimates of efficiences. The linear regression model is glu = 56.831-0.875npreg+0.104bp+0.263skin+0.796bmi+0.764age.

Problem2

- **Nonlinearity Check**

```
summary(fit1)$r.squared
```

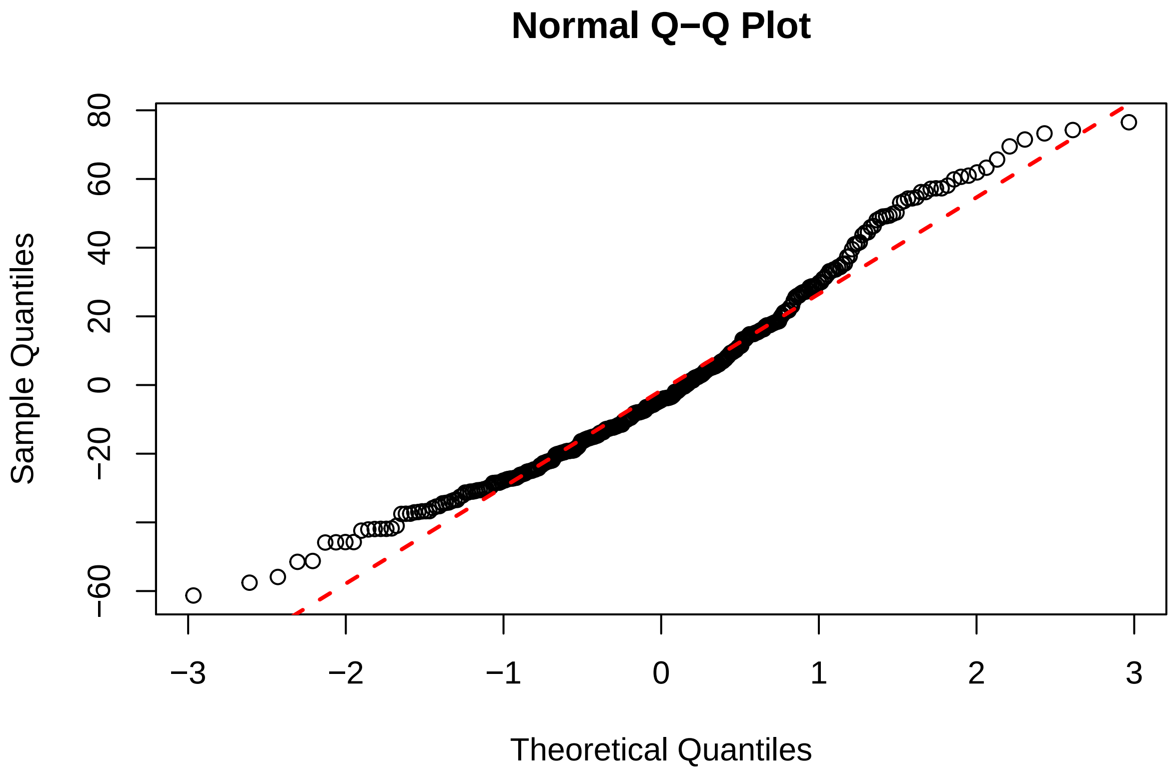
```
## [1] 0.1338
```

Since the R-squared is only 0.1338033, really small, which means there are a lot of variance is not explained by the model. It suggests the fitted model suffers from the lack of fit.

Also, from the plot of residuals against fitted value, we could see the relation between fitted value and residual is curvilinear. Thus, the lineariry assumption is not true.

- **Normality Check**

```
qqnorm(fit1$residuals)
qqline(fit1$residuals,col = 2,lwd=2,lty=2)
```



The normal probability plot with a concave-upward shape shows the distribution of error term is left-skewed. The assumption of normality is invalid.

```
st <- shapiro.test(fit1$residuals)
st
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.9703, p-value = 2.532e-06
```

```
st$p.value
```

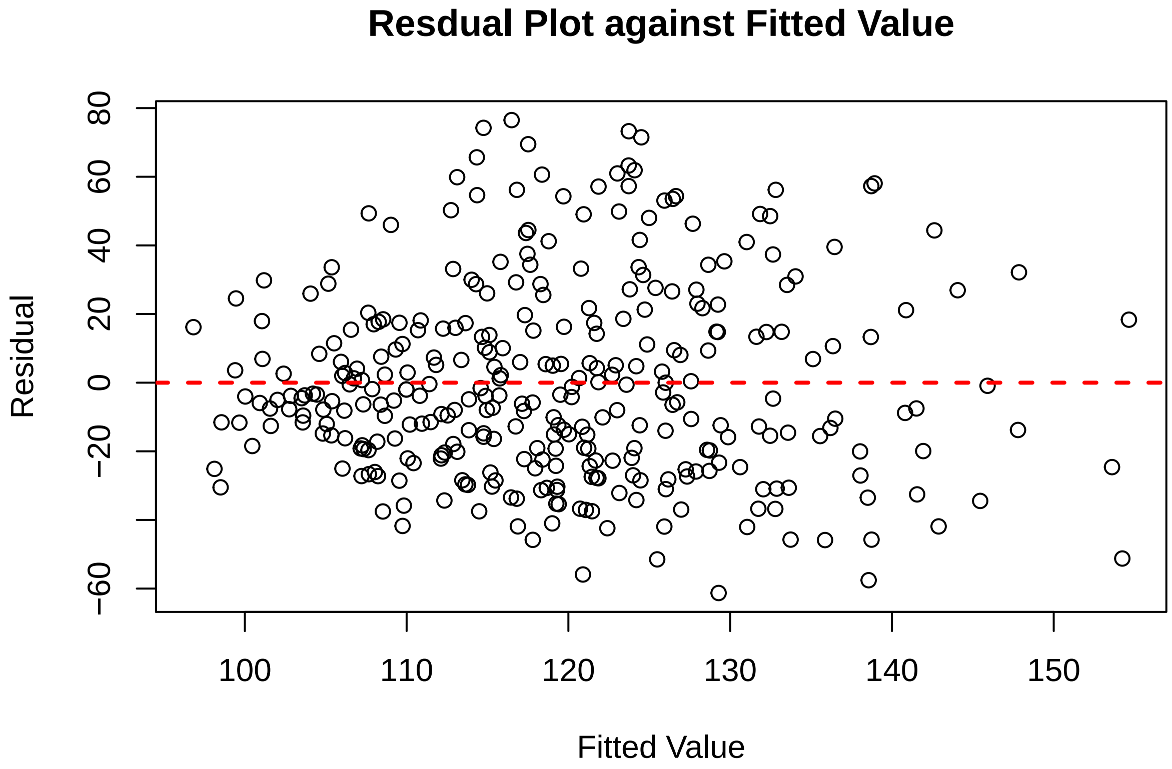
```
## [1] 2.532e-06
```

Since the p-value from Shapiro-Wilk test is significantly small, we could conclude that the sample deviates from normality.

- **Homoscedasticity Check**

The plot of the residuals against the fitted values is also helpful to examine the homoscedasicity of the error term.

```
plot(fit1$fitted.values,fit1$residuals,xlab='Fitted Value',ylab='Residual',main="Resdual Plot against F
abline(h=0,col = 2,lwd=2,lty=2)
```



Again, the residuals fall around 0, showing no tendencies and certian pattern, which means the variance of the error terms is constant.

- **Uncorrelated Error Check**

```
require(lmtest)
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 3.0.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.0.3
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dw <- dwtest(fit1)
dw
```

```
##
##  Durbin-Watson test
##
## data:  fit1
## DW = 1.938, p-value = 0.2847
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dw$p.value
```

```
## [1] 0.2847
```

Since the p-value from Durbin-watson test is greater than 0.05, there is no evidence to reject the null hypothesis that there is a correlation within the error term. The assumption of uncorrelated error is valid.

- **Outliers**

```
###Examine outlying Y observations
n = nrow(data)
elist = fit1$resi
p = 6
SSE = sum(elist^2)
X = cbind(1,data$npreg,data$bp,data$skin,data$bmi,data$age)
hlist = diag(X%*%solve(t(X)%*%X)%*%t(X))
tlist = elist*((n-p-1)/(SSE*(1-hlist)-elist^2))^(1/2)
max(abs(tlist))
```

```
## [1] 2.707
```

```
which(abs(tlist)==max(abs(tlist)))
```

```
## 177
## 177
```

```
qt(0.9975,n-p-1)
```

```
## [1] 2.826
```

Using Bonferrono simultaneous test procedure with a family significance level 0.01, we have t(0.9975,325)=2.826329. Since the absolute value of largest absolute studentized deleted residual is 2.706896, smaller than t(0.9975,325)=2.826329, we conclude that the case 177 is not an outlier.

```
###Identifying outlying X observations with Hat Matrix Leverage Values
2*p/n
```

```
## [1] 0.03614
```

```
which(hlist > 2*p/n)
```

```
##  [1]   5   8  12  18  21  41  43  57  72  79  92 107 141 158 196 198 203
## [18] 211 217 232 249 262 287 291 292 320 330
```

From the result, we could identify the outliers of X.

- **Influential Points**

```
lmi <- lm.influence(fit1)
lms <- summary(fit1)
e <- resid(fit1)
s <- lms$sigma
si <- lmi$sigma
xxi <- diag(lms$cov.unscaled)
h <- lmi$hat
bi <- coef(fit1)-t(coef(lmi))
dfbetas <- bi/t(si%o%xxi^0.5)
stand.resid <- e/(si*(1-h)^0.5)
DFFITS <- h^0.5*e/(si*(1-h))
which(abs(stand.resid)>2*sqrt(p/n))
```

```
##   1   2   3   4   5   6   7   8  12  13  15  16  17  18  19  20  21  22
##   1   2   3   4   5   6   7   8  12  13  15  16  17  18  19  20  21  22
##  23  26  27  28  29  30  31  33  34  35  36  38  39  41  42  43  45  46
##  23  26  27  28  29  30  31  33  34  35  36  38  39  41  42  43  45  46
##  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
##  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
##  66  67  68  69  70  71  72  73  74  75  76  77  78  79  82  86  87  88
##  66  67  68  69  70  71  72  73  74  75  76  77  78  79  82  86  87  88
##  89  90  91  94  95  96  97  99 100 101 102 103 104 105 106 107 110 113
##  89  90  91  94  95  96  97  99 100 101 102 103 104 105 106 107 110 113
## 114 115 117 119 123 124 125 127 128 129 130 131 132 134 135 137 138 139
## 114 115 117 119 123 124 125 127 128 129 130 131 132 134 135 137 138 139
## 140 141 143 145 146 148 149 150 151 152 153 156 157 158 159 160 161 162
## 140 141 143 145 146 148 149 150 151 152 153 156 157 158 159 160 161 162
## 163 164 165 166 168 169 170 172 173 174 175 176 177 178 179 180 181 183
## 163 164 165 166 168 169 170 172 173 174 175 176 177 178 179 180 181 183
## 184 185 186 187 188 189 190 191 192 193 194 195 196 198 199 201 202 203
## 184 185 186 187 188 189 190 191 192 193 194 195 196 198 199 201 202 203
## 205 206 207 208 209 210 211 212 214 215 216 217 218 219 220 222 223 224
## 205 206 207 208 209 210 211 212 214 215 216 217 218 219 220 222 223 224
## 225 226 227 228 230 231 233 234 235 236 238 239 240 241 242 243 244 246
## 225 226 227 228 230 231 233 234 235 236 238 239 240 241 242 243 244 246
## 247 248 249 250 251 253 254 257 258 259 260 262 264 265 266 267 268 269
## 247 248 249 250 251 253 254 257 258 259 260 262 264 265 266 267 268 269
## 270 271 272 273 274 276 277 278 280 281 282 284 286 289 290 291 292 293
## 270 271 272 273 274 276 277 278 280 281 282 284 286 289 290 291 292 293
## 294 295 297 298 299 300 301 304 305 306 307 308 309 311 312 314 315 316
## 294 295 297 298 299 300 301 304 305 306 307 308 309 311 312 314 315 316
## 317 320 321 322 323 326 328 329 330 331 332
## 317 320 321 322 323 326 328 329 330 331 332
```

We could use DFFITS to check the influential points. The index of cases show above.

Problem 3. The remedial measures in case of violations of any of the underlying assumptions

- **Lack of fit:** 1) Simple trasformations, e.g, take log; 2)Non-linear model; 3)Other predictors.
- **Non-constancy:** 1) Transformation; 2) Build variance structure in to model: WLS.
- **Non-Normality:** 1) Transformation; 2)Robust regression methods.
- **Correlated Errors:**1) Transformation: Cochrane-Crutt Procedure. 2) Use models that incorporate the correlattion structure: Generalized Estimating Equations
- **Multicollinearity:** Ridge regression
- **Influential Cases:** Robust regression

From problem 2, we know the nonlinearity, non-normality, and influential cases exist. We could use proposal of remedial measures above to fix problems.