Stat 151 Fall 2015

Homework 5 Solutions

December 6, 2015

1. The first 9 plots are shown in Figure 1

```
bfat = read.table('bodyfat_corrected.txt',header = TRUE)

# fitting linear model and getting diagnostics
lmodel = lm(BODYFAT~ AGE + WEIGHT + HEIGHT + THIGH,data = bfat)
h = influence(lmodel)$hat
predres = lmodel$residuals/(1-h)
stdpredres = rstudent(lmodel)

# making plots
p = NULL
p[[1]] = qplot(lmodel$fitted.values, lmodel$residuals, cex = 0.3, xlab = 'Fitted Values', ylab = 'Residu
p[[2]] = qplot(lmodel$fitted.values, rstandard(lmodel), cex = 0.3, xlab = 'Fitted Values', ylab = 'Stand
p[[3]] = qplot(rstandard(lmodel), lmodel$residuals, cex = 0.3, xlab = 'Standardized Residuals', ylab = '
p[[4]] = qplot(lmodel$fitted.values, predres, cex = 0.3, xlab = 'Fitted Values', ylab = 'Predicted Resid
p[[5]] = qplot(lmodel$residuals, predres, cex = 0.3, xlab = 'Residuals', ylab = 'Predicted Residuals') +
p[[6]] = qplot(h, lmodel$residuals, cex = 0.3, xlab = 'Leverage', ylab = 'Residuals') + scale_size_conti
p[[7]] = qplot(stdpredres, predres,cex = 0.3, xlab = 'Studentized Residuals', ylab = 'Predicted Residual
p[[8]] = qplot(stdpredres, rstandard(lmodel), cex = 0.3, xlab = 'Studentized Residuals', ylab = 'Standar
p[[9]] = qplot(1:nrow(bfat), cooks.distance(lmodel), cex = 0.3, xlab = 'Observation ID', ylab = 'Cook\'s

multiplot(plotlist = p, cols = 3)
```

Note that the plots are column major order. In case of confusion double check the axes labels.

Comments: From the plots in Figure 1, we see some potential outliers and influential points. Let us use the `which` function in `R` to fish out problematic observations.

In the first plot we see at least 4 observations removed from the data cloud, three to the right of the plot and one at the very top. The following checks give us the IDs of these observations, `which(abs(lmodel$residuals) > 15)` and `which(abs(lmodel$fitted.values) > 35)` which combined give us the IDs `39 41 42 216` where the fit is potentially poor.

The second plot establishes this along with indicating potential minor problems at IDs `36 207`

The third plot suggests that the residuals and standardized residuals vary significantly (all others lie almost on a straight line) at ID 42. This might indicate an unusual leverage at this point.

The fourth point corroborates our previous findings. The fifth plot further tags IDs `39 42` as having unusual leverage.

The sixth plot lets us pick out high leverage observations quite easily. Recalling that $\sum_i h_i = p$ so that the average leverage is p/n, we decide to tag any observation with leverage more than $2p/n$ as high leverage. The following points are then tagged: `15 29 39 41 42 79 108 147 169 216 239 242` . Note that this list is quite different from the list of potential outliers. This is because leverage calculation does not look at the responses. Four observations, with IDs `39 41 42 216` appears on both lists, meaning both high leverage and a poor fit.

From the seventh and eighth plots, IDs `39 42` are again tagged for being away from the straight line and from the cloud. The last plot tags the same two observations as having unusually high Cook's distance.

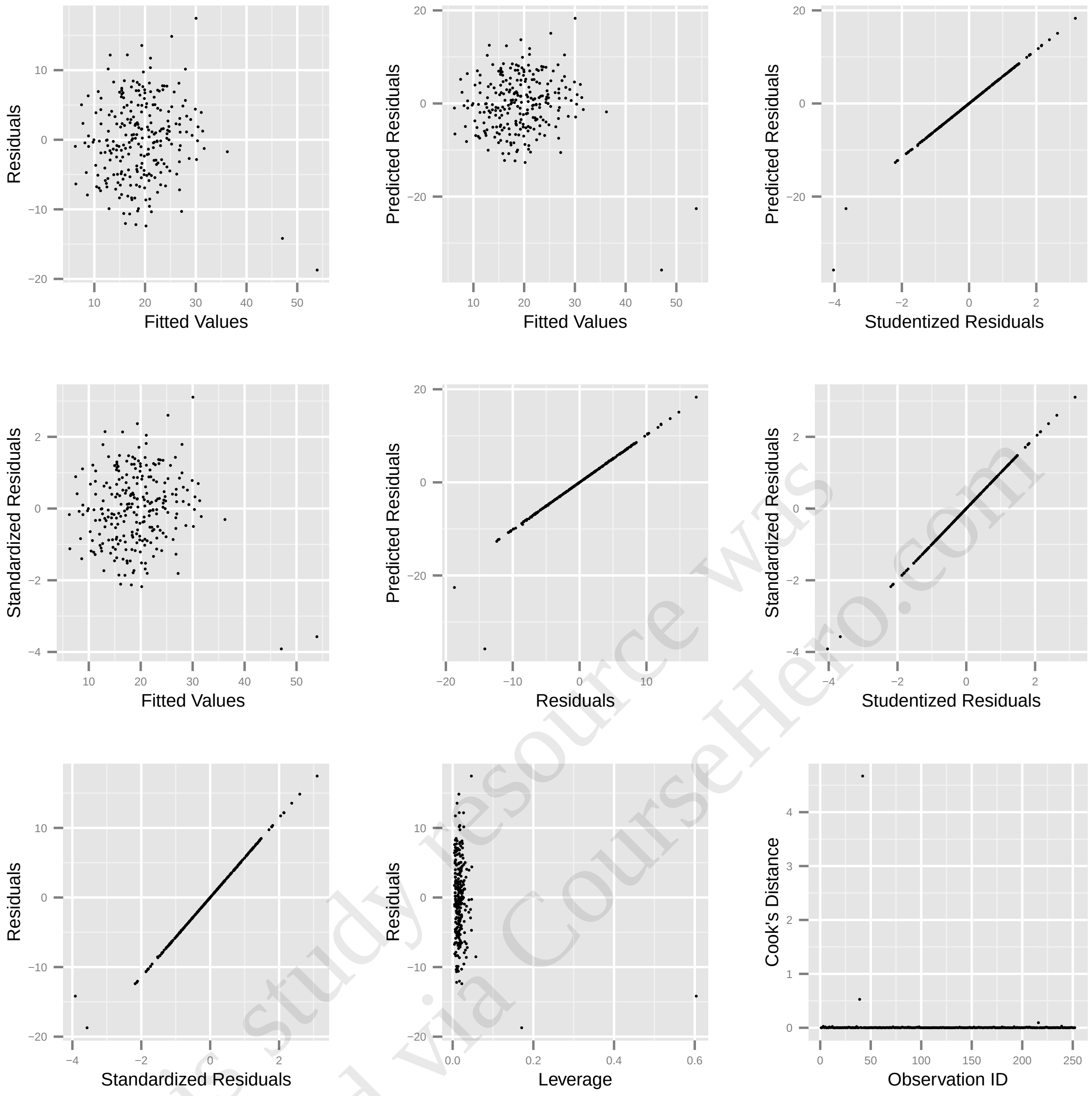


Figure 1: diagnostic plots for BODYFAT dataset

Based on these findings, observations 39 42 stand out as outliers. Other potential outliers are 36 41 207 216 and other high influence points are 15 29 79 108 147 169 239 242

The second plot of p-values of testing whether the i^{th} observation is an outlier based on the studentizerd residuals is shown in Figure 2 with a red line marking 0.05.

```
# compute p-values for each ID based on stdpredres
pvals = sapply(stdpredres,function(t) 1-pt(t,235))
qplot(1:nrow(bfat),pvals,size = 1,, xlab = 'Observation ID', ylab = 'p-values') + geom_bar(stat = 'ident
```

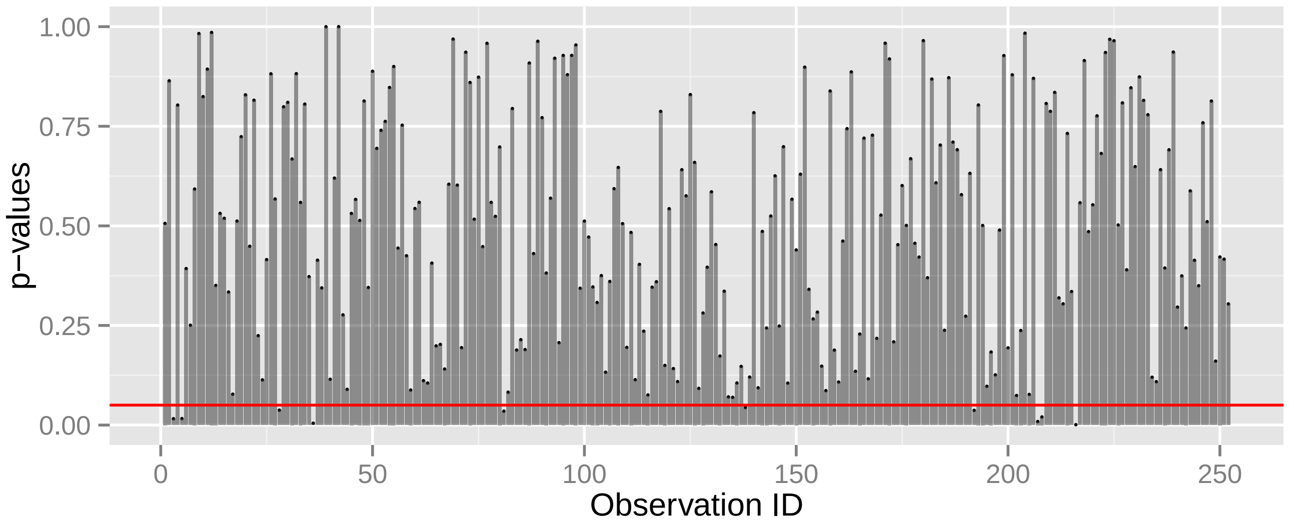


Figure 2: p-values for each observation

```
# use vertical lines, red horizontal line at 0.05
```

This marks observations `3 5 28 36 81 138 192 207 208 216` as outliers. But this is incorrect. Each of the tests (for individual observations) have a 5% chance of tagging that observation as outlier even when it is not. As we are doing 252 such tests, one for each observation, it is expected that we can see as many as 13 tags even in the absence of any true outliers. (In this experiment we see 10)

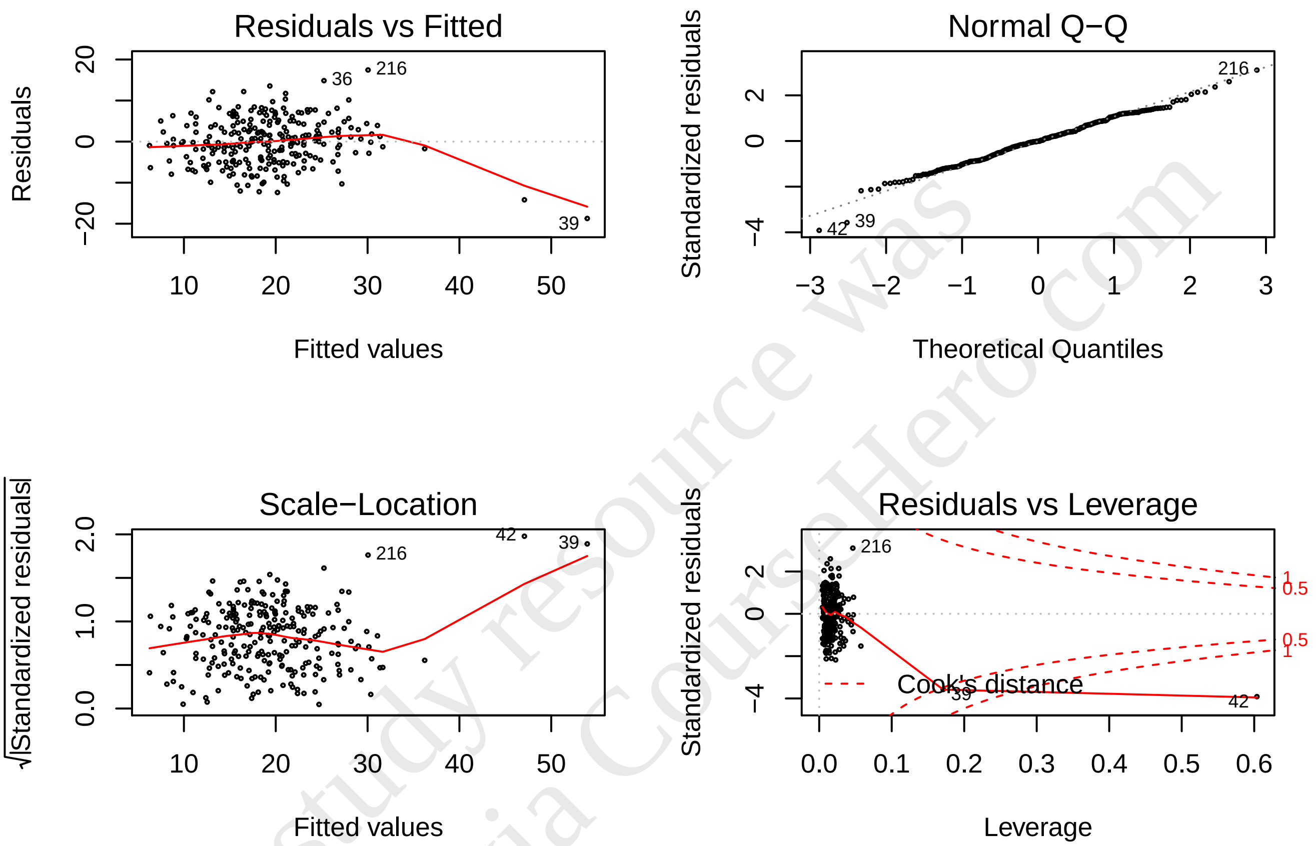
One can correct this issue by tagging an observation as an outlier only if its p-value is smaller than 0.05/252. This correction is in general overly conservative. This does not give us any positives on this dataset.

Based on all the indications we have seen above, we can prescribe that observations `39 42` be removed from the dataset. The summary of this new analysis is given below

```
bfat2 = bfat[-c(39,42),]
summary(lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH, data = bfat2))

##
## Call:
## lm(formula = BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH, data = bfat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4982  -3.7381  -0.0034   3.7581  12.0943
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.82844   13.74245   3.117  0.00205 **
## AGE          0.16101    0.03164   5.089 7.18e-07 ***
## WEIGHT       0.21150    0.03020   7.003 2.39e-11 ***
## HEIGHT      -1.18281    0.16753  -7.060 1.70e-11 ***
## THIGH        0.24418    0.15252   1.601  0.11068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.365 on 245 degrees of freedom
## Multiple R-squared:  0.5883,Adjusted R-squared:  0.5816
## F-statistic: 87.54 on 4 and 245 DF,  p-value: < 2.2e-16
```

2.
```
g = lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + THIGH, data = bfat)
par(mfrow = c(2, 2))
plot(g, cex = 0.3)
```



The first plot shows the residuals against the fitted values. We expect the residuals to form a cloud around the x-axis and constant spread. The red line is the loess fit of the residuals with the fitted values, which is also expected to be close to the x-axis if the linear model assumptions are verified. In this plot, we see the existence of a few points near the right hand margin of the plot which pull the loess line away from the x-axis, indicating possible outliers

The second plot is the Q-Q plot of the standardized residuals. If the errors in the linear model are in fact gaussian we expect the plot to closely follow the straight line $y = x$. We see here that the left tail of the standardized residuals is heavier than normal.

The third plot is a check for heteroscedasticity and again the expected shape is similar to that in the first plot. We see some points which towards the right margin of the plot which move the red loess line away from the x-axis

The fourth plot shows three measurements at once - standardized residuals, leverage and cook's distance. The interpretation of Cook's distance in this plot is through the red contours. Regions towards the right and away from the x-axis have higher Cook's distance. This plot indicates the interaction between influence and outlyingness of the observations. We see a few observations that have unusually high Cook's distance.

3. (a)

$$\begin{aligned}\hat{y}_i - \hat{y}_{i(i)} &= (y_i - \hat{y}_{i(i)}) - (y_i - \hat{y}_i)\\ &= \hat{e}_{i(i)} - \hat{e}_i\\ &= \hat{e}_i(\tfrac{1}{1-h_{ii}} - 1)\\ &= \hat{e}_i\tfrac{h_{ii}}{1-h_{ii}}\end{aligned}$$

(b) Using the fact that $\hat{e}_i \sim N(0, (1-h_{ii})\sigma^2)$, we get

$$\hat{y}_i - \hat{y}_{i(i)} \sim N(0, \sigma^2 \frac{h_{ii}^2}{1-h_{ii}})$$

(c) For this part we need to find an unbiased estimator of σ^2 which is independent of $\hat{y}_i - \hat{y}_{i(i)}$ or equivalently of $\hat{e}_i$. It can be verified that

$$\frac{RSS_{(i)}}{n-p-2}$$

is an unbiased estimator of σ^2, indeed it is the natural estimator of σ^2 in the smaller dataset not containing the i^{th} observation. From Lecture Eighteen, this estimator is also independent $\hat{e}_i$.

4. Let $y_{(j)}$ denote the residual upon regressing y on $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ and $x_{(j)}$ denote the residual upon regressing x_j on $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$. We will use the following fact

The regression coefficient upon regressing $y_{(j)}$ upon $x_{(j)}$ is the same as $\hat{\beta}_j$ (the regression coefficient of x_j in the multiple regression of y on $x_1, \ldots, x_p$)

So the slopes we see in the partial regression plots correspond to the estimated coefficients we see in the summary. This lets us conclude that,

Plot 1 corresponds to x_2

Plot 2 corresponds to x_3

Plot 3 corresponds to x_1

The RSS of any partial regression is same as the RSS of the original regression, `431.1933`

5. (a)

```
bfat = bfat[,-1]
# backward selection by p-values
flag = 1
bfdat = bfat
while(flag){
  lmi = summary(lm(BODYFAT ~. , data = bfdat))
  pvals = lmi$coefficients[-1,4]
  weakest = which.max(pvals)
  if (pvals[weakest] >= 0.05){
    bfdat = bfdat[,-(weakest + 1)]
  } else {
    flag = 0
  }
}
model1 = names(bfdat)[-1]

# Forward selection using p-values
flag = 1
selected = NULL
outside = 2:ncol(bfat)
s = 'BODYFAT ~ 1'
namevec = names(bfat)
while(flag){
  small = lm(s, data = bfat)
  pvals = array(0, dim = length(outside))
  for (i in 1:length(outside)){
    big = lm(paste(s, ' + ', namevec[outside[i]]), data = bfat)
    pvals[i] = anova(small,big)[2,6]
  }

  best = which.min(pvals)
  if (pvals[best] < 0.05){
    s = paste(s, ' + ', namevec[outside[best]])
    selected = append(selected, namevec[outside[best]])
    outside = outside[-best]
  } else {
    flag = 0
  }
}
model2 = selected


# Forward selection by adjusted R^2
flag = 1
selected = NULL
outside = 2:ncol(bfat)
s = 'BODYFAT ~ 1'
namevec = names(bfat)
while(flag){

  smallr2 = summary(lm(s, data = bfat))$adj.r.squared
  bigr2 = array(0, dim = length(outside))
  for (i in 1:length(outside)){
    bigr2[i] = summary(lm(paste(s, ' + ', namevec[outside[i]]), data = bfat))$adj.r.squared
  }

  best = which.max(bigr2)
  if (bigr2[best] > smallr2){
    s = paste(s, ' + ', namevec[outside[best]])
    selected = append(selected, namevec[outside[best]])
    outside = outside[-best]
  } else {
    flag = 0
  }
}
model3 = selected


# Forward selection by AIC
flag = 1
selected = NULL
outside = 2:ncol(bfat)
s = 'BODYFAT ~ 1'
namevec = names(bfat)
while(flag){

  smallaic = extractAIC(lm(s, data = bfat))[2]
  bigaic = array(0, dim = length(outside))
  for (i in 1:length(outside)){
    bigaic[i] = extractAIC(lm(paste(s, ' + ', namevec[outside[i]]), data = bfat))[2]
  }

  best = which.min(bigaic)
  if (bigaic[best] < smallaic){
    s = paste(s, ' + ', namevec[outside[best]])
    selected = append(selected, namevec[outside[best]])
    outside = outside[-best]
  } else {
    flag = 0
  }
}
model4 = selected

# Forward selection by BIC
flag = 1
selected = NULL
outside = 2:ncol(bfat)
s = 'BODYFAT ~ 1'
namevec = names(bfat)
while(flag){

  smallbic = extractAIC(lm(s, data = bfat), k = log(nrow(bfat)))[2]
  bigbic = array(0, dim = length(outside))
  for (i in 1:length(outside)){
    bigbic[i] = extractAIC(lm(paste(s, ' + ', namevec[outside[i]]), data = bfat), k = log(nrow(bfat)
  }

  best = which.min(bigbic)
  if (bigbic[best] < smallbic){
    s = paste(s, ' + ', namevec[outside[best]])
    selected = append(selected, namevec[outside[best]])
    outside = outside[-best]
  } else {
    flag = 0
  }
}
model5 = selected

# For normal likelihood, Mallow's C_p is equivalent to AIC
model6 = model4

cat('Model selected by backward selection by individual p-values is\n', model1)
```

```
## Model selected by backward selection by individual p-values is
##  WEIGHT ABDOMEN FOREARM WRIST
```

```
cat('Model selected by forward selection by p-values is\n', model2)
```

```
## Model selected by forward selection by p-values is
##  ABDOMEN WEIGHT WRIST FOREARM
```

```
cat('Model selected by backward selection by adjusted R-squared is\n', model3)
```

```
## Model selected by backward selection by adjusted R-squared is
##  ABDOMEN WEIGHT WRIST FOREARM NECK AGE THIGH HIP BICEPS

cat('Model selected by backward selection by AIC is\n', model4)

## Model selected by backward selection by AIC is
##  ABDOMEN WEIGHT WRIST FOREARM NECK AGE THIGH HIP

cat('Model selected by forward selection by BIC is\n', model5)

## Model selected by forward selection by BIC is
##  ABDOMEN WEIGHT WRIST FOREARM

cat('Model selected by forward selection by Mallow\'s C_p is\n', model6)

## Model selected by forward selection by Mallow's C_p is
##  ABDOMEN WEIGHT WRIST FOREARM NECK AGE THIGH HIP
```

(b) Now performing cross validation on the selected models

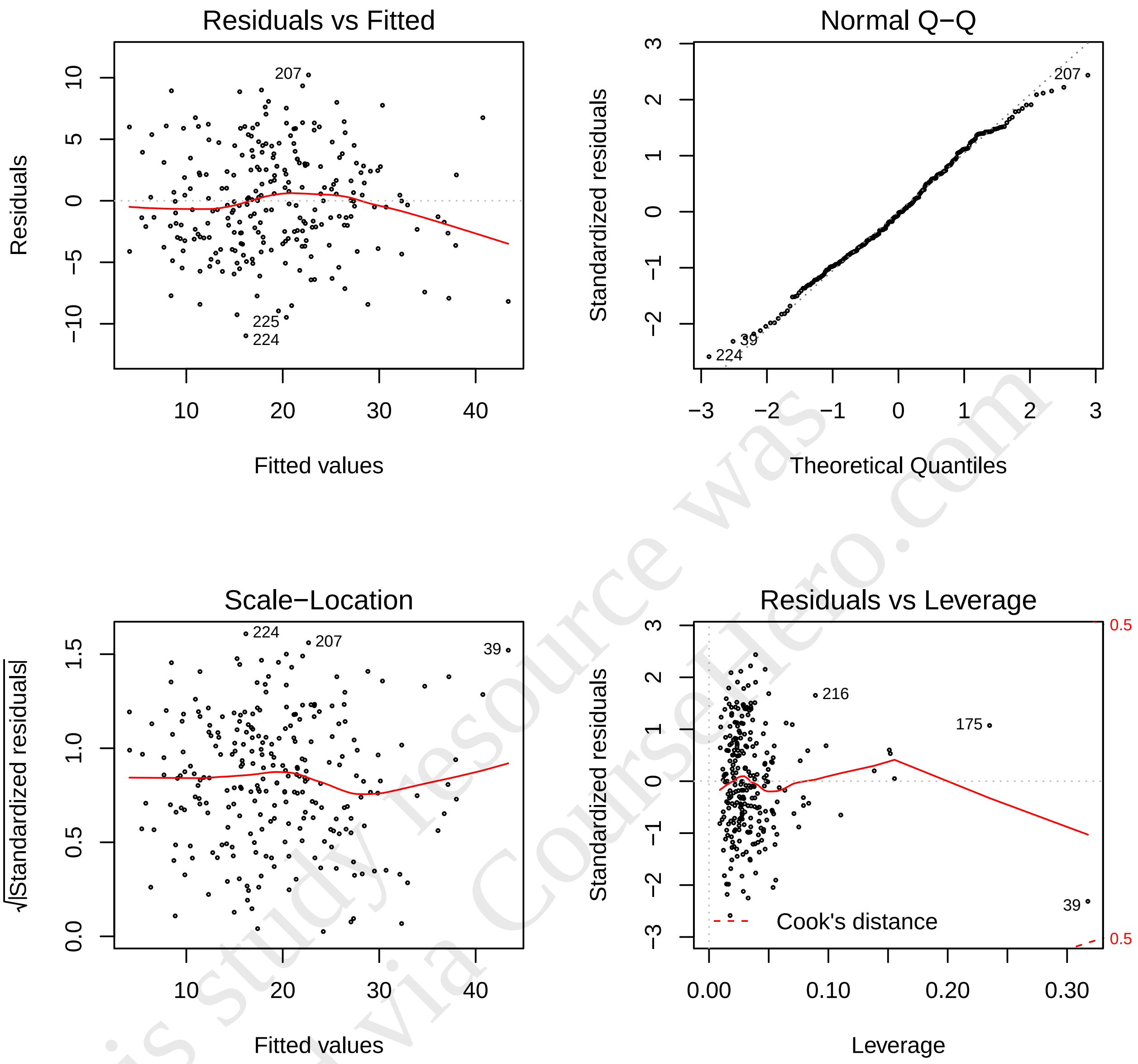
```
bfatlist = NULL
bfatlist[[1]] = bfat[,c('BODYFAT',model1)]
bfatlist[[2]] = bfat[,c('BODYFAT',model2)]
bfatlist[[3]] = bfat[,c('BODYFAT',model3)]
bfatlist[[4]] = bfat[,c('BODYFAT',model4)]
bfatlist[[5]] = bfat[,c('BODYFAT',model5)]
bfatlist[[6]] = bfat[,c('BODYFAT',model6)]

nfolds = 10
n = nrow(bfat)
foldid = sample(nfolds, n, replace = T)
pred.err = matrix(0, nfolds, 6)

for (k in 1:nfolds){
  inset = subset(1:n, foldid != k)
  outset = subset(1:n, foldid == k)
  for (j in 1:6){
    fit = lm(BODYFAT ~. , data = bfatlist[[j]][inset,])
    preds = predict(fit, bfatlist[[j]][outset,-1])
    pred.err[k,j] = sqrt(mean((preds - bfatlist[[j]][outset,1])^2))
  }
}

pe.vec = apply(pred.err,2,mean)
cat('The average prediction errors of the six models are \n',pe.vec)

## The average prediction errors of the six models are
##  4.424874 4.424874 4.373924 4.362751 4.424874 4.362751

cat('The best prediction error is given by model number ',which.min(pe.vec))

## The best prediction error is given by model number  4
```

Which seems to imply that AIC selected the best model. It should be noted that the average prediction errors are very similar and any difference can possibly be attributed to the randomness in fold selection.

(c) Regression diagnostics on the AIC model shows

```
bfbest = bfatlist[[4]]
g = lm(BODYFAT ~ ., data = bfbest)
par(mfrow = c(2, 2))
```

```r
plot(g, cex = 0.3)
```



```r
hg = influence(g)$hat
cdg = cooks.distance(g)
```

The first plot does not reveal any deviations from the assumptions of the linear model. The second plot reveals that the standardized residuals are lighter tailed than normal. The third and fourth plot tag observation `39` as a potential outlier, with leverage of `0.32` and Cook's Distance of `0.28`.

Some other potential high influence points are `36 39 159 175 206`

Compared to the diagnostics in question 1, these numbers are very clean. This inspection does not reveal any points that can be justifiably deleted.