STAT 151A HW3 Solutions for #2 and #3

Billy Fang

2

Recall the definitions

$$r_i := \frac{\hat{e}_i}{\sqrt{\mathrm{RSS}/(n-p-1)}\sqrt{1-h_i}}, \qquad t_i := \frac{\hat{e}_i}{\sqrt{\mathrm{RSS}_{[i]}/(n-p-2)}\sqrt{1-h_i}}$$

Then,

$$\begin{aligned}
\left(\frac{r_i}{t_i}\right)^2 &= \frac{\mathrm{RSS}_{[i]}/(n-p-2)}{\mathrm{RSS}/(n-p-1)} \\
&= \frac{n-p-1}{n-p-2}\left(1 - \frac{\hat{e}_i^2/(1-h_i)}{\mathrm{RSS}}\right) && \mathrm{RSS}_{[i]} = \mathrm{RSS} - \frac{\hat{e}_i^2}{1-h_i} \\
&= \frac{1}{n-p-2}\left(n-p-1 - \frac{\hat{e}_i^2}{(\mathrm{RSS}/(n-p-1))\cdot(1-h_i)}\right) \\
&= \frac{n-p-1-r_i^2}{n-p-2}.
\end{aligned}$$

Some rearranging yields the desired equality.

3

```
bodyfat <- read.csv("bodyfat.csv")
n <- dim(bodyfat)[1]
p <- 4
mod <- lm(bodyfat ~ Age + Weight + Height + Thigh, data=bodyfat)
res <- resid(mod)
fit <- fitted(mod)
sigmahat <- summary(mod)$sigma
X <- as.matrix(cbind(1, bodyfat[,c("Age", "Weight", "Height", "Thigh")]))
lev <- hat(X)
```

Although you were not asked to do this, let us look at a plot of the leverage values.

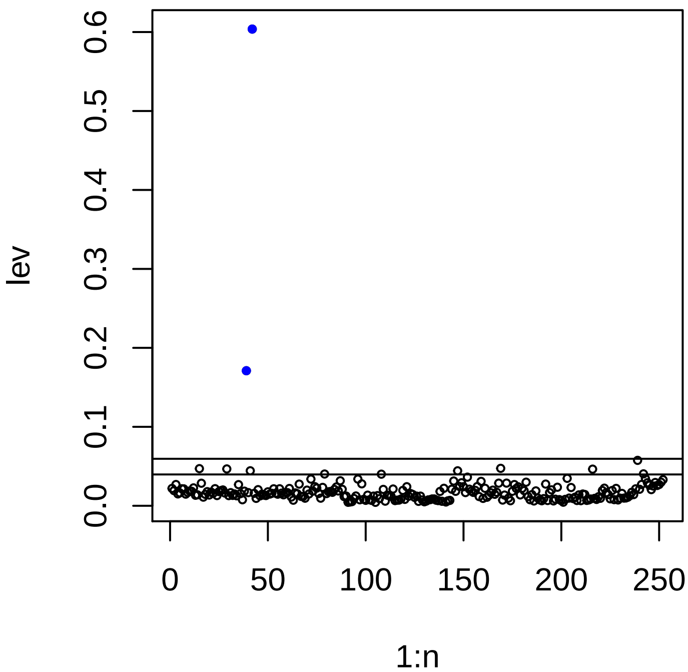
```
plot(1:n, lev, cex=0.5)
hbar <- (p+1)/n
abline(h = 2 * hbar)
abline(h = 3 * hbar)
lev.sorted <- sort(lev, decreasing=T, index.return=T)
n.big <- length(which(lev > 3 * hbar))
idx <- lev.sorted$ix[1:n.big]
idx
```

```
## [1] 42 39
```

```r
lev[idx]
```

```
## [1] 0.6037373 0.1710321
```

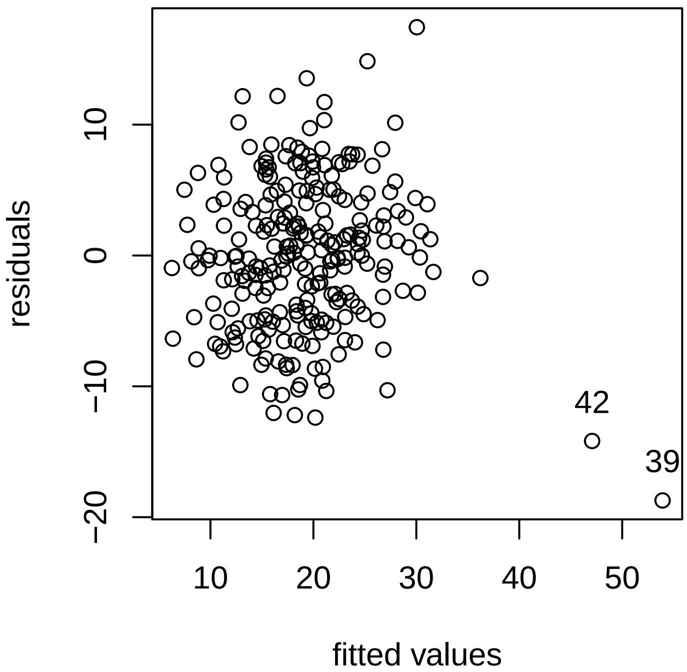
```r
points(idx, lev[idx], pch=19, cex=0.5, col='blue')
```



We have two high-leverage points: subjects 42 and 39.

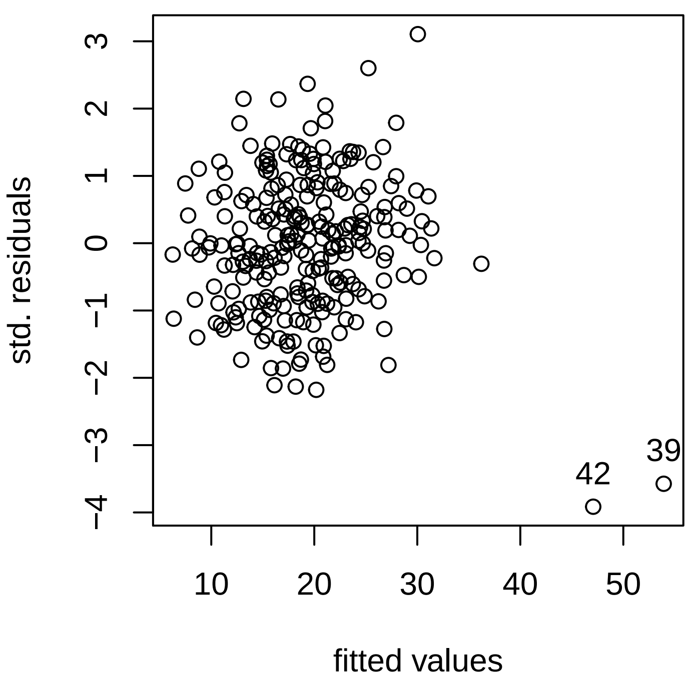
a)

```r
plot(fit, res, xlab="fitted values", ylab="residuals")
text(fit[42], res[42]+3, "42")
text(fit[39], res[39]+3, "39")
```



b)

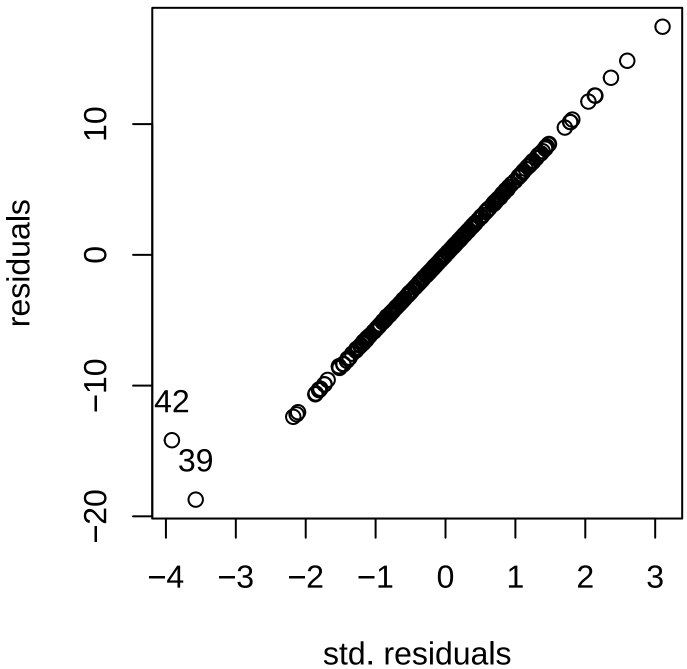
```
res.std <- res / (sigmahat * sqrt(1 - lev))
plot(fit, res.std, xlab="fitted values", ylab="std. residuals")
text(fit[42], res.std[42]+0.5, "42")
text(fit[39], res.std[39]+0.5, "39")
```




```
# plot(fit, rstandard(mod))
```

c)

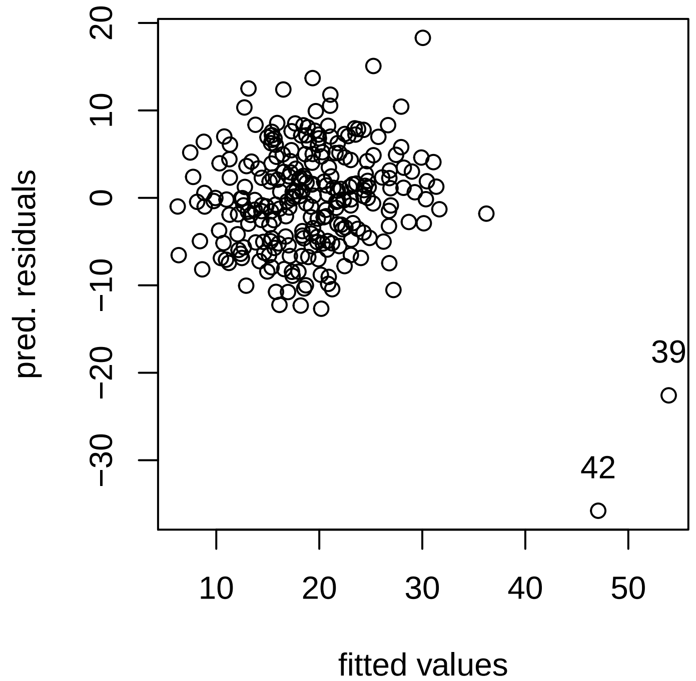
```
plot(res.std, res, xlab="std. residuals", ylab="residuals")
text(res.std[42], res[42]+3, "42")
text(res.std[39], res[39]+3, "39")
```




```
# plot(rstandard(mod), res)
```

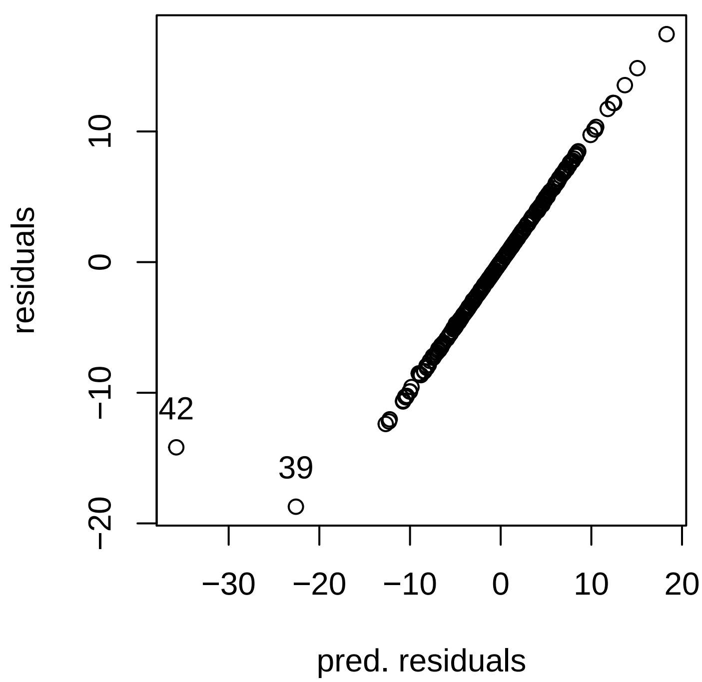
d)

```
res.pred <- res / (1 - lev)
plot(fit, res.pred, xlab="fitted values", ylab="pred. residuals")
text(fit[42], res.pred[42]+5, "42")
text(fit[39], res.pred[39]+5, "39")
```



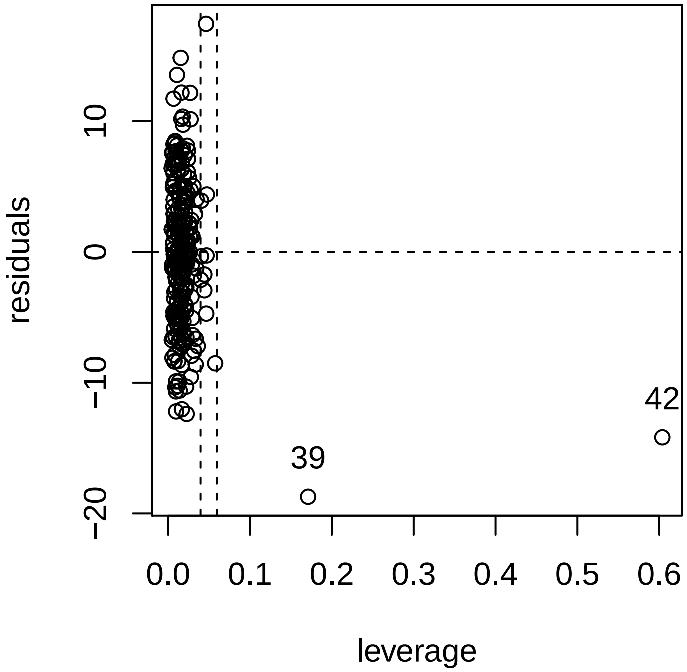
e)

```r
plot(res.pred, res, xlab="pred. residuals", ylab="residuals")
text(res.pred[42], res[42]+3, "42")
text(res.pred[39], res[39]+3, "39")
```



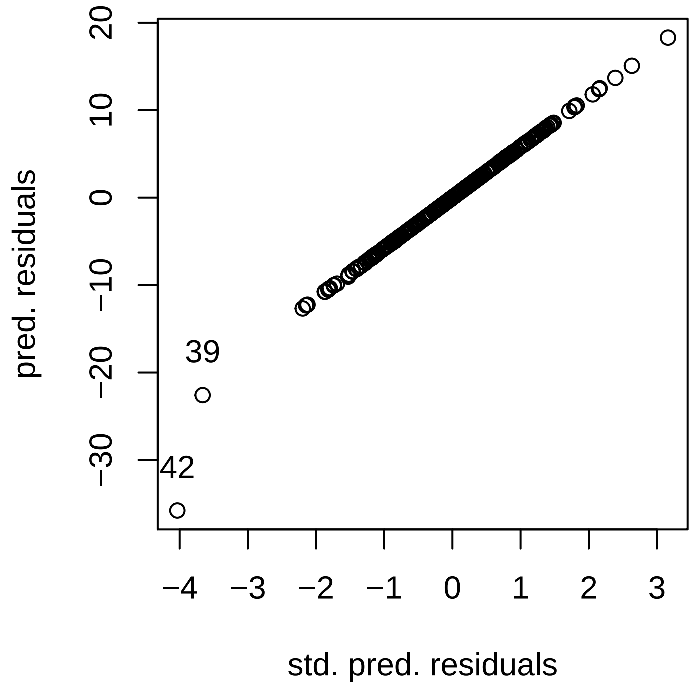
f)

```
plot(lev, res, xlab="leverage", ylab="residuals")
abline(v = 2 * hbar, lty=2)
abline(v = 3 * hbar, lty=2)
text(lev[42], res[42]+3, "42")
text(lev[39], res[39]+3, "39")
abline(h = 0, lty=2)
```



g)

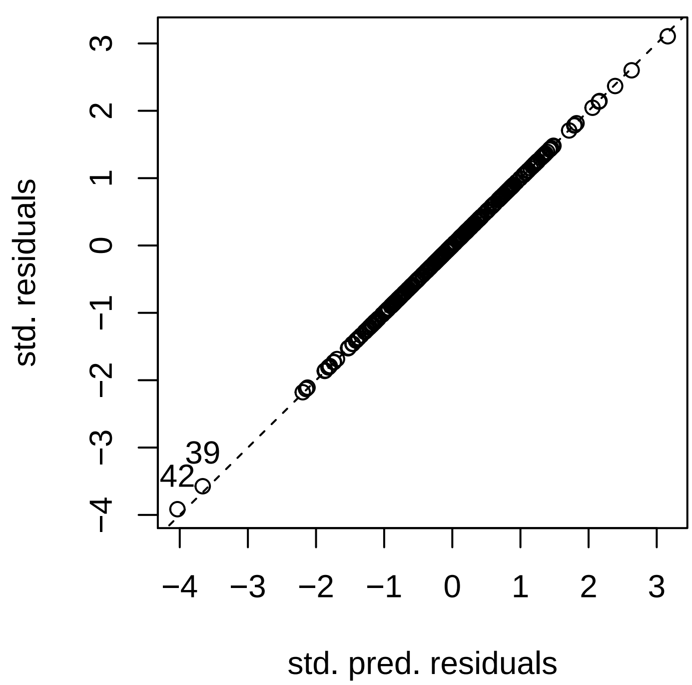
```
res.stdpred <- res.std * sqrt((n - p - 2) / (n - p - 1 - res.std^2))
plot(res.stdpred, res.pred, xlab="std. pred. residuals", ylab="pred. residuals")
text(res.stdpred[42], res.pred[42]+5, "42")
text(res.stdpred[39], res.pred[39]+5, "39")
```




```
# plot(rstudent(mod), res.pred)
```

h)

```
plot(res.stdpred, res.std, xlab="std. pred. residuals", ylab="std. residuals")
text(res.stdpred[42], res.std[42]+0.5, "42")
text(res.stdpred[39], res.std[39]+0.5, "39")
abline(a=0, b=1, lty=2)
```



```
# plot(rstudent(mod), rstandard(mod))
```

i)

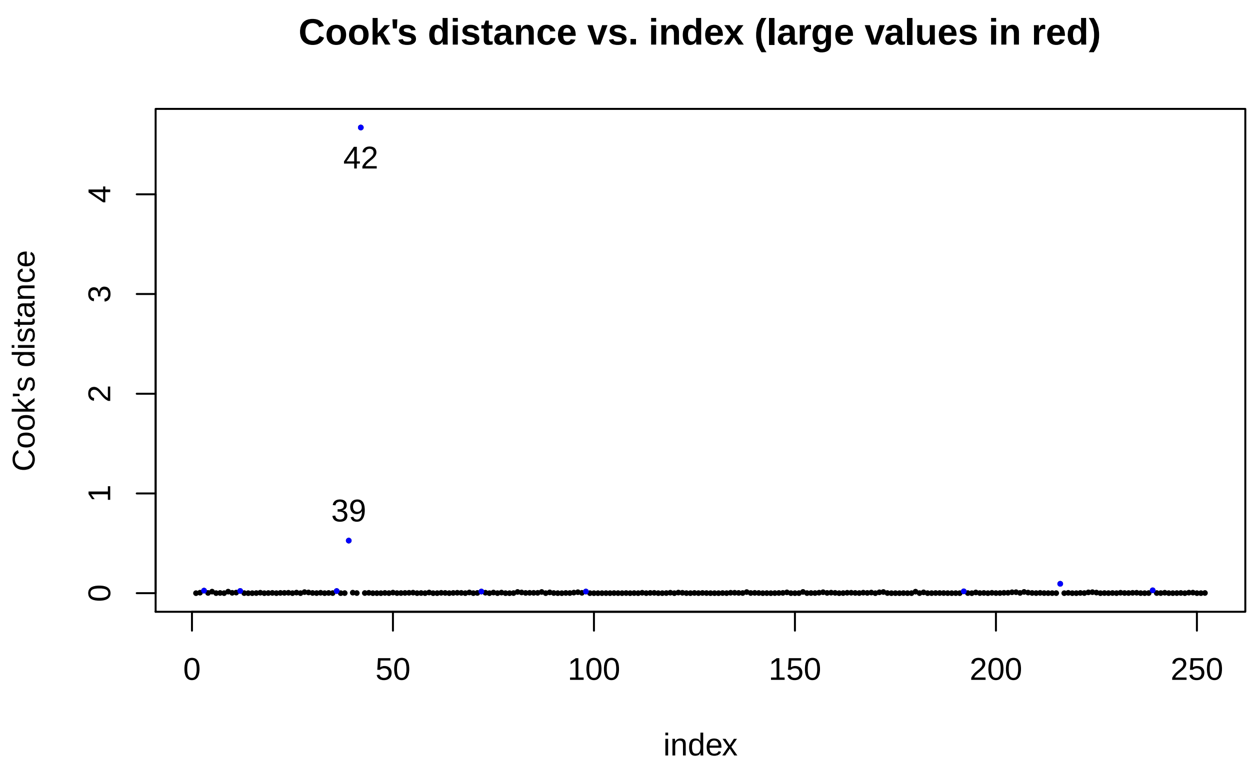
```
cook <- res.std^2 / (p + 1) * lev / (1 - lev)
plot(1:n, cook, pch=16, cex=0.4, xlab="index", ylab="Cook's distance",
     main="Cook's distance vs. index (large values in red)")
cook.sorted <- sort(cook, decreasing=T, index.return=T)
cook.cutoff <- 4 / (n - p - 1) # see page 282
num.big <- length(which(cook > cook.cutoff))
idx <- cook.sorted$ix[1:num.big]
cook[idx]
```

```
##         42         39        216        239          3         12
## 4.66969177 0.52711418 0.09388593 0.02839649 0.02547257 0.02201873
##         36        192         98         72
## 0.02113771 0.01807339 0.01623765 0.01620523
```

```
num.big
```

```
## [1] 10
```

```
points(idx, cook[idx], pch=16, cex=0.4, col='blue')
text(42, cook[42]-0.3, "42")
text(39, cook[39]+0.3, "39")
```



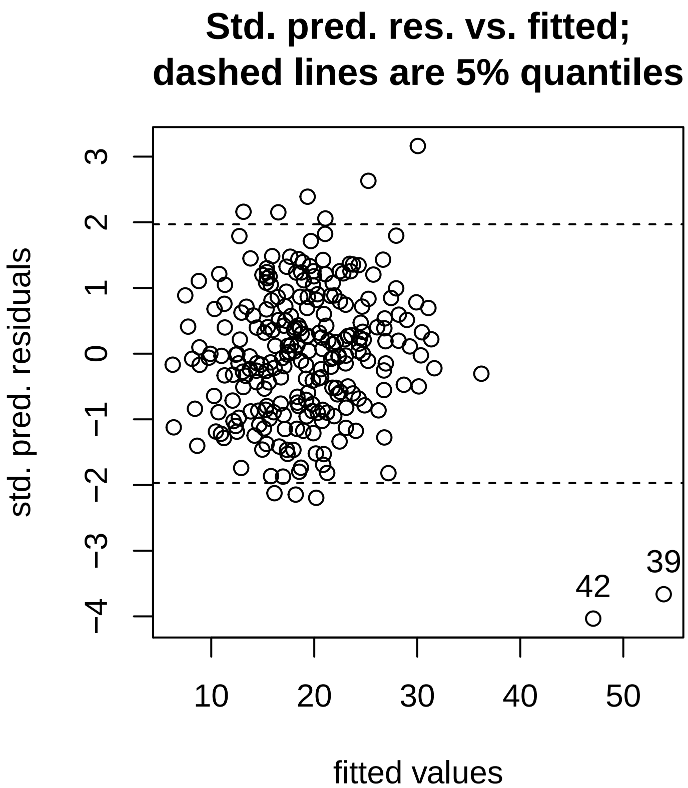
The cutoff chosen above is $\frac{4}{n-p-1}$ as suggested by page 282 of the textbook.

j)

The above plots (a)-(g) suggest that the two high-leverage points 42 and 39 are outliers. They already stand out in the residual plot (a), but standardization (b) makes this even more obvious. The various other plots support this reasoning as well.

We include the following plot to further support this claim.

```r
plot(fit, res.stdpred, xlab="fitted values", ylab="std. pred. residuals",
     main="Std. pred. res. vs. fitted;\ndashed lines are 5% quantiles")
q <- qt(0.975, n - p - 1)
abline(h=-q, lty=2)
abline(h=q, lty=2)
text(fit[42], res.stdpred[42]+0.5, "42")
text(fit[39], res.stdpred[39]+0.5, "39")
```



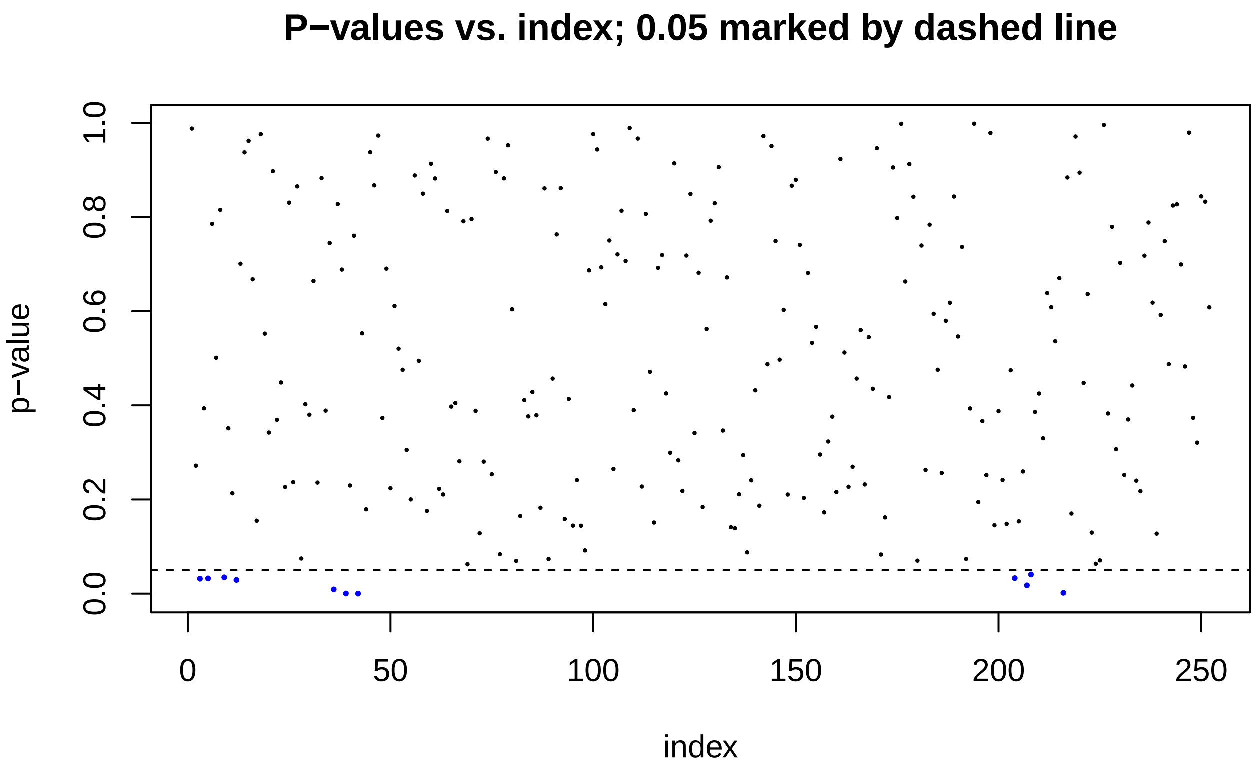
We see that after studentization, these two points have much larger residuals than the other datapoints. The two lines are the quantiles for a 5% t-test, and the eleven points beyond these lines are the eleven blue points in the p-value plot in the next part.

From our work in part i), we see that there are a few influential points. The most influential are the two high-leverage points 42 and 39, followed by 8 other points above the recommended cutoff, but these 8 other points have much smaller influence than the two high-leverage points.

k)

```r
pval <- 2 * (1 - pt(abs(res.stdpred), n - p - 2))
plot(1:n, pval, pch=16, cex=0.3, xlab="index", ylab="p-value",
     main="P-values vs. index; 0.05 marked by dashed line")
```

```r
abline(h=0.05, lty=2)
#abline(h=0, lty=3)
#abline(h=1, lty=3)
pval.sorted <- sort(pval, index.return=T)
num.small <- length(which(pval < 0.05))
idx <- pval.sorted$ix[1:num.small]
points(idx, pval[idx], pch=16, cex=0.4, col='blue')
```




```r
pval[idx]
```

```
##           42           39          216           36          207
## 7.325343e-05 3.054138e-04 1.764147e-03 9.029060e-03 1.760332e-02
##           12            3            5          204            9
## 2.909586e-02 3.169823e-02 3.237542e-02 3.285382e-02 3.461962e-02
##          208
## 4.062107e-02
```

```r
num.small
```

```
## [1] 11
```

It does not make sense to rule all of these high p-values as outliers at a 5% level since these are separate t-tests, and some of them may be large purely by chance. We would need to do a Bonferroni correction to properly reject the largest residual at a 5% level.

```r
num.small <- length(which(pval < 0.05 / n))
idx <- pval.sorted$ix[1:num.small]
pval[idx]
```

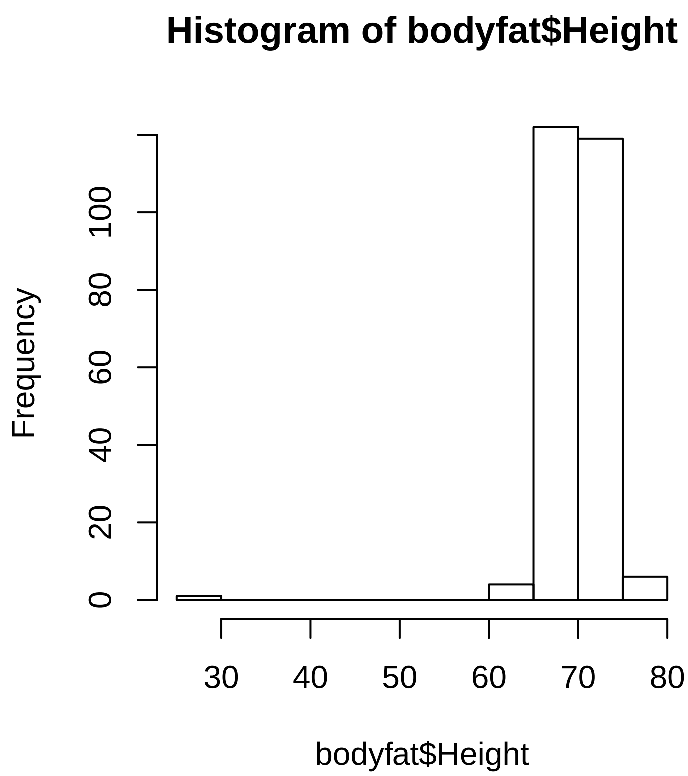
```
##           42
## 7.325343e-05
```

From here, we see that the test only rejects subject 42.

l)

We should be careful when removing data. The analysis above suggests that subjects 42 and 39 have something interesting, but we should look more closely at the data to understand what exactly is going on.

```
hist(bodyfat$Height)
```




```
bodyfat[42,]
```

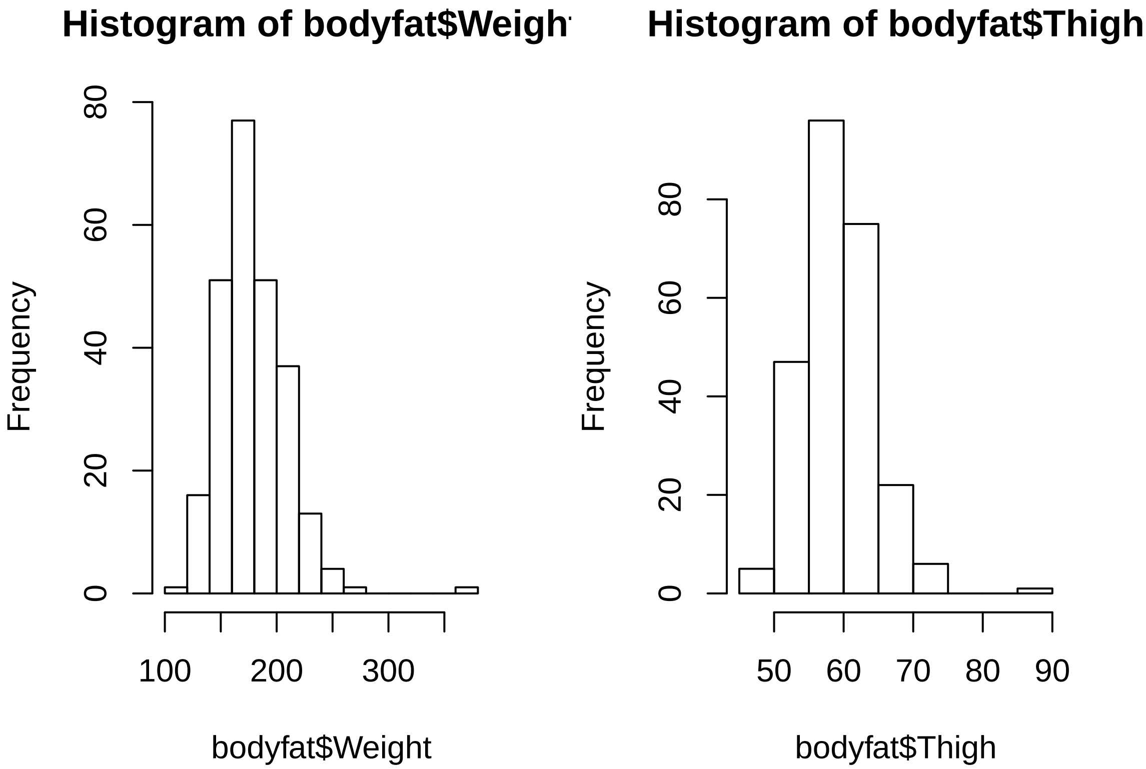
```
##    Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee
## 42   1.025    32.9  44    205   29.5 36.6   106   104.3 115.5  70.6 42.5
##    Ankle Biceps Forearm Wrist
## 42  23.7   33.6    28.7  17.4
```

A quick glance at the data shows that subject 42 has an extremely low height of 29.5 inches (2 ft 5.5 in, or 74.93 cm), which is probably the main source of this data point's high leverage. This is likely a recording error, since this subject weights 205 pounds and is 44 years old. Since we do not have a way to correct this error, we should probably remove this point as the above analysis suggests.

What about subject 39?

```
hist(bodyfat$Weight)
hist(bodyfat$Thigh)
bodyfat[39,]
```

```
##    Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee
## 39  1.0202    35.2  46 363.15  72.25 51.2 136.2   148.1 147.7  87.3 49.1
##    Ankle Biceps Forearm Wrist
## 39  29.6     45      29  21.4
```



We can see that subject 39 has high leverage due to his relatively high weight and thigh measurement that is beyond the range of everyone else in the study; he is over 100 pounds heavier than the next heaviest person in the study. Regarding whether we should remove this datapoint from the dataset or not, one can argue either way: one one hand one might feel that this subject is unusual and our model should only account for males with measurements closer to the rest of the data, while on the other hand one might feel that we have no reason to exclude this subject from the analysis, since it may give some insight into improving our model to account for males with larger measurements. We do not explore this further.

We list the summaries for the original model, the model fitted after removing subject 42, and the model fitted after removing both subjects 42 and 39 below.

```r
summary(mod)
```

```
##
## Call:
## lm(formula = bodyfat ~ Age + Weight + Height + Thigh, data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.722  -4.283  -0.055   4.061  17.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.27488   11.12642  -0.204   0.8382
## Age          0.20517    0.03274   6.267 1.63e-09 ***
## Weight       0.13417    0.02952   4.545 8.59e-06 ***
## Height      -0.49810    0.11313  -4.403 1.59e-05 ***
## Thigh        0.38970    0.16142   2.414   0.0165 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.753 on 247 degrees of freedom
## Multiple R-squared:  0.5349, Adjusted R-squared:  0.5274
## F-statistic: 71.03 on 4 and 247 DF,  p-value: < 2.2e-16
```

```r
mod2 <- lm(bodyfat ~ Age + Weight + Height + Thigh, data=bodyfat[-42,])
summary(mod2)
```

```
##
## Call:
## lm(formula = bodyfat ~ Age + Weight + Height + Thigh, data = bodyfat[-42,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.2729  -3.7828  -0.0947   3.9254  13.0096
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.86048   14.18928   2.457   0.0147 *
## Age          0.17168    0.03284   5.228 3.66e-07 ***
## Weight       0.17257    0.03019   5.717 3.13e-08 ***
## Height      -1.02550    0.17072  -6.007 6.77e-09 ***
## Thigh        0.29942    0.15824   1.892   0.0596 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.583 on 246 degrees of freedom
## Multiple R-squared:  0.559,  Adjusted R-squared:  0.5519
## F-statistic: 77.96 on 4 and 246 DF,  p-value: < 2.2e-16
```

```r
mod3 <- lm(bodyfat ~ Age + Weight + Height + Thigh, data=bodyfat[-c(39,42),])
summary(mod3)
```

```
##
## Call:
## lm(formula = bodyfat ~ Age + Weight + Height + Thigh, data = bodyfat[-c(39,
##     42), ])
##
## 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
```