Lab11

2024-11-24

Questions

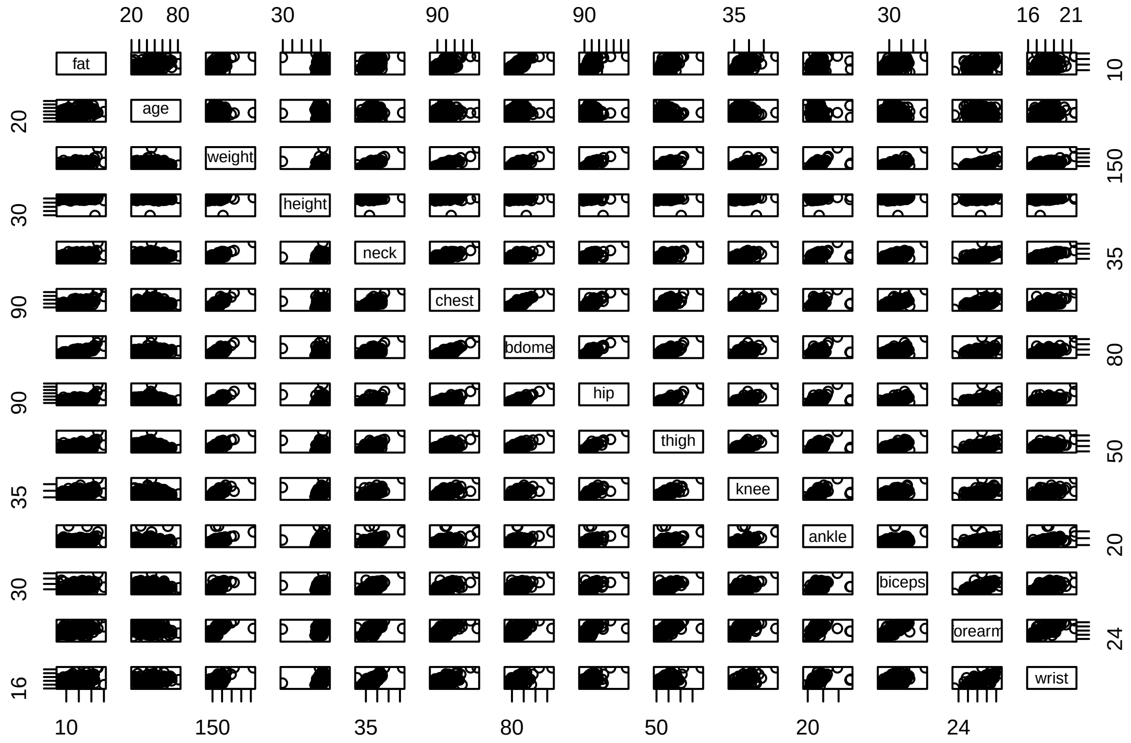
- Q3/4 (c): Do we add a note of the "intercept" for these problems?
- Q6 (a): Units of "fat" variable; also, is it appropriate to use "conditional mean" since we're comparing "intercepts"?
- Q7, Q8

```
library(readr)
bodyfat <- read_table("bodyfat.txt", col_types = cols(density = col_skip()))
```

1.

Summarize your findings from examining the pairwise scatterplots and correlation matrix.

```
pairs(bodyfat)
```



```r
cor(bodyfat)
```

```
##                fat         age     weight      height       neck      chest
## fat      1.0000000  0.31758778  0.5463420 -0.17031967 0.40941003 0.65923536
## age      0.3175878  1.00000000 -0.0425167 -0.10033083 0.08045261 0.09479893
## weight   0.5463420 -0.04251670  1.0000000  0.22050342 0.80751029 0.89280857
## height  -0.1703197 -0.10033083  0.2205034  1.00000000 0.22961509 0.09911784
## neck     0.4094100  0.08045261  0.8075103  0.22961509 1.00000000 0.75820034
## chest    0.6592354  0.09479893  0.8928086  0.09911784 0.75820034 1.00000000
## abdomen  0.7902383  0.20025327  0.8760832  0.02843775 0.71184385 0.90275741
## hip      0.5737411 -0.07857168  0.9316951  0.04036149 0.70035253 0.82998129
## thigh    0.5349748 -0.22543254  0.8534585 -0.01299705 0.67617454 0.74729063
## knee     0.5005588 -0.02544728  0.8432788  0.13281129 0.63932948 0.72295657
## ankle    0.2184638 -0.08824517  0.5437907  0.18502990 0.39781675 0.42624020
## biceps   0.4167874 -0.06633859  0.7632409  0.13058391 0.67999435 0.71198471
## forearm  0.3479230 -0.11600518  0.6845928  0.21326139 0.65805982 0.66070774
## wrist    0.2855033  0.20151964  0.7157644  0.31919880 0.72947461 0.64793598
##            abdomen         hip       thigh        knee       ankle      biceps
## fat     0.79023826  0.57374109  0.53497482  0.50055877  0.21846384  0.41678744
## age     0.20025327 -0.07857168 -0.22543254 -0.02544728 -0.08824517 -0.06633859
## weight  0.87608320  0.93169513  0.85345850  0.84327878  0.54379067  0.76324091
## height  0.02843775  0.04036149 -0.01299705  0.13281129  0.18502990  0.13058391
## neck    0.71184385  0.70035253  0.67617454  0.63932948  0.39781675  0.67999435
## chest   0.90275741  0.82998129  0.74729063  0.72295657  0.42624020  0.71198471
## abdomen 1.00000000  0.86905574  0.76695941  0.74145618  0.38503752  0.64259286
## hip     0.86905574  1.00000000  0.89285888  0.81152623  0.48197659  0.70530315
## thigh   0.76695941  0.89285888  1.00000000  0.79138923  0.43238025  0.73746222
## knee    0.74145618  0.81152623  0.79138923  1.00000000  0.53021195  0.65031775
## ankle   0.38503752  0.48197659  0.43238025  0.53021195  1.00000000  0.39724029
## biceps  0.64259286  0.70530315  0.73746222  0.65031775  0.39724029  1.00000000
## forearm 0.52824720  0.57201543  0.62858826  0.59848341  0.36918092  0.71841055
## wrist   0.58646517  0.58308984  0.50828740  0.62175951  0.48665981  0.59193260
##            forearm     wrist
## fat      0.3479230 0.2855033
## age     -0.1160052 0.2015196
## weight   0.6845928 0.7157644
## height   0.2132614 0.3191988
## neck     0.6580598 0.7294746
## chest    0.6607077 0.6479360
## abdomen  0.5282472 0.5864652
## hip      0.5720154 0.5830898
## thigh    0.6285883 0.5082874
## knee     0.5984834 0.6217595
## ankle    0.3691809 0.4866598
## biceps   0.7184105 0.5919326
## forearm  1.0000000 0.6871657
## wrist    0.6871657 1.0000000
```

For our response variable (Y) of “fat” (bodyfat), all of our potential explanatory variables have at least a magnitude of 0.15, and a majority (all but one) appear to be positively linearly related to “fat”. The strongest correlation (largest in magnitude) is approximately 0.79 for the explanatory variable abdomen, such that we overall have reason to believe that a linear fit between our explanatory variables with the response would be appropriate at first glance.

However, when we compare the correlations and pairwise plots between explanatory variables we also observe potentially strong linear relationships (not always, but a fair number, particularly for explanatory variables of hip, thigh, and knee). Generally speaking, we observe fairly strong correlations between parts of the body close in proximity to one another, such as the three variables mentioned previously.

Despite having some potential concerns about multicollinearity (to be explored in later questions), we also see that a number of explanatory variables have a very weak linear relationship between one another, such as ankle and age having only a magnitude of correlation of 0.08, or other body combinations with the “age” explanatory variable. This means we have potential reason to believe including more than one explanatory variable could be helpful for our model without excessive risk of multicollinearity.

That is all to say: I believe we are tasked with identifying “strong” correlation by whether |r| > 0.7. To that end the following are “strong” (or “significant”) correlations: (1) Fat (response variable) and abdomen, for just comparisons amongst explanatory variables, there are many, including (2) weight and neck, chest, abdomen, hip, thigh, knee, biceps, and wrist, (3) neck and chest, abdomen, hip, and wrist, (4) chest and weight, neck, abdomen, hip, thigh, knee, and biceps, (5) abdomen and fat, weight, neck, chest, hip, thigh, and knee, (6) hip and weight, chest, abdomen, thigh, knee, and biceps, and more.

2.

Discuss whether the VIFs indicate any explanatory variables exhibiting extreme multicollinearity.

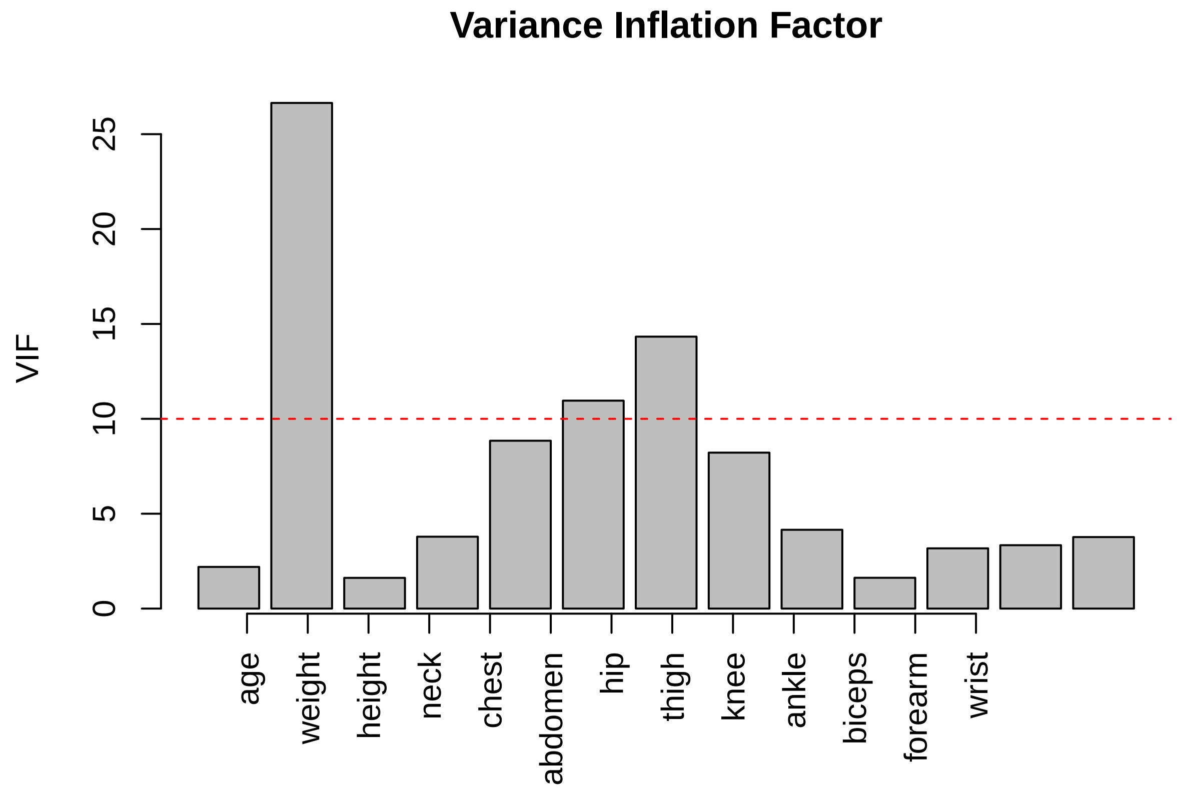
```
full.fat <- lm(fat~., data=bodyfat)
summary(full.fat)
```

```
##
## Call:
## lm(formula = fat ~ ., data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4856 -2.7408 -0.4312  2.6985  8.7373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.924231  21.243400  -1.220  0.22475
## age           0.110678   0.042329   2.615  0.01009 *
## weight       -0.115065   0.062928  -1.829  0.06997 .
## height       -0.082485   0.102958  -0.801  0.42464
## neck         -0.411145   0.274755  -1.496  0.13720
## chest         0.008897   0.130496   0.068  0.94576
## abdomen       0.946729   0.111393   8.499 6.47e-14 ***
## hip          -0.279862   0.183271  -1.527  0.12940
## thigh         0.331539   0.197092   1.682  0.09516 .
## knee          0.154684   0.331130   0.467  0.64125
## ankle         0.392981   0.236996   1.658  0.09992 .
## biceps       -0.042424   0.209941  -0.202  0.84020
## forearm       0.992959   0.358416   2.770  0.00650 **
## wrist        -2.199682   0.755444  -2.912  0.00429 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.095 on 119 degrees of freedom
## Multiple R-squared:  0.7717, Adjusted R-squared:  0.7468
## F-statistic: 30.94 on 13 and 119 DF,  p-value: < 2.2e-16
```

```
library(car)
```

```
## Loading required package: carData
```

```
vif.fat <- vif(full.fat)
barplot(vif.fat, xaxt = "n", main = "Variance Inflation Factor", ylab = "VIF")
# barplot(vif.fat)
abline(h=10, col="red", lty=2)
axis(1, at = seq_along(vif.fat), labels = names(vif.fat), las = 2)
```



Given the prompt, “VIF values larger than 10 indicate severe multicollinearity”, we observe three explanatory variables having a VIF value larger than 10 and have reason to suspect potential issues of multicollinearity. The explanatory variables of note are “weight”, “abdomen”, and “hip”.

3.

Summarize the backward elimination method of model selection by providing:

```
back.fat <- step(full.fat, direction="backward")
```

```
## Start:  AIC=388.2
## fat ~ age + weight + height + neck + chest + abdomen + hip +
##     thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - chest    1      0.08 1995.5 386.21
## - biceps   1      0.68 1996.1 386.25
## - knee     1      3.66 1999.1 386.44
## - height   1     10.76 2006.2 386.92
## <none>                 1995.4 388.20
## - neck     1     37.55 2033.0 388.68
## - hip      1     39.10 2034.5 388.78
## - ankle    1     46.11 2041.6 389.24
## - thigh    1     47.45 2042.9 389.33
## - weight   1     56.07 2051.5 389.89
## - age      1    114.64 2110.1 393.63
## - forearm  1    128.70 2124.1 394.51
## - wrist    1    142.17 2137.6 395.35
## - abdomen  1   1211.23 3206.7 449.29
##
## Step:  AIC=386.21
## fat ~ age + weight + height + neck + abdomen + hip + thigh +
##     knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - biceps   1      0.65 1996.2 384.25
## - knee     1      3.59 1999.1 384.44
## - height   1     11.35 2006.9 384.96
## <none>                 1995.5 386.21
## - neck     1     37.48 2033.0 386.68
## - hip      1     40.02 2035.5 386.85
## - ankle    1     46.12 2041.6 387.24
## - thigh    1     47.98 2043.5 387.37
## - weight   1     60.86 2056.4 388.20
## - age      1    114.70 2110.2 391.64
## - forearm  1    139.38 2134.9 393.19
## - wrist    1    142.75 2138.3 393.39
## - abdomen  1   1712.56 3708.1 466.61
##
## Step:  AIC=384.25
## fat ~ age + weight + height + neck + abdomen + hip + thigh +
##     knee + ankle + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - knee     1      3.81 2000.0 382.50
## - height   1     11.27 2007.4 383.00
## <none>                 1996.2 384.25
## - neck     1     38.27 2034.4 384.77
## - hip      1     39.78 2036.0 384.87
## - ankle    1     46.10 2042.3 385.29
## - thigh    1     47.92 2044.1 385.40
## - weight   1     64.72 2060.9 386.49
## - age      1    114.27 2110.4 389.65
## - wrist    1    142.16 2138.3 391.40
## - forearm  1    150.39 2146.6 391.91
## - abdomen  1   1732.70 3728.9 465.36
##
## Step:  AIC=382.5
## fat ~ age + weight + height + neck + abdomen + hip + thigh +
##     ankle + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - height   1     11.03 2011.0 381.23
## <none>                 2000.0 382.50
## - hip      1     39.88 2039.9 383.13
## - neck     1     44.19 2044.2 383.41
## - ankle    1     51.59 2051.6 383.89
## - thigh    1     60.66 2060.6 384.48
## - weight   1     60.91 2060.9 384.49
## - age      1    123.02 2123.0 388.44
## - wrist    1    139.16 2139.1 389.45
## - forearm  1    151.56 2151.5 390.22
## - abdomen  1   1728.96 3728.9 463.36
##
## Step:  AIC=381.23
## fat ~ age + weight + neck + abdomen + hip + thigh + ankle + forearm +
##     wrist
##
##           Df Sum of Sq    RSS    AIC
## <none>                 2011.0 381.23
## - hip      1     32.30 2043.3 381.35
## - neck     1     44.89 2055.9 382.17
## - ankle    1     53.67 2064.7 382.74
## - thigh    1     78.71 2089.7 384.34
## - weight   1     99.52 2110.5 385.66
## - age      1    141.13 2152.1 388.25
## - wrist    1    152.74 2163.7 388.97
## - forearm  1    157.48 2168.5 389.26
## - abdomen  1   1798.41 3809.4 464.20
```

```r
summary(back.fat)
```

```
##
## Call:
## lm(formula = fat ~ age + weight + neck + abdomen + hip + thigh +
##     ankle + forearm + wrist, data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5547 -2.8437 -0.2409  2.6936  8.8349
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30.91894   15.50171  -1.995  0.04830 *
## age           0.11847    0.04032   2.938  0.00394 **
## weight       -0.12734    0.05162  -2.467  0.01500 *
## neck         -0.44046    0.26581  -1.657  0.10006
## abdomen       0.96082    0.09161  10.488  < 2e-16 ***
## hip          -0.24520    0.17445  -1.406  0.16236
## thigh         0.38262    0.17438   2.194  0.03010 *
## ankle         0.41844    0.23095   1.812  0.07245 .
## forearm       0.99644    0.32106   3.104  0.00237 **
## wrist        -2.24539    0.73463  -3.056  0.00275 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.043 on 123 degrees of freedom
## Multiple R-squared:  0.7699, Adjusted R-squared:  0.7531
## F-statistic: 45.73 on 9 and 123 DF,  p-value: < 2.2e-16
```

(a)

An ordered list of which variable was removed from the model at each step;

Step 1: “chest” removed Step 2: “bicep” removed Step 3: “knee” removed Step 4: “height” removed

(b)

A list of which variables remained in the final model;

Variables kept in final model: “age”, “weight”, “neck”, “abdomen”, “hip”, “thigh”, “ankle”, “foearm”, and “wrist”.

(c)

A summary of the partial regression coefficients effects tests for the final model.

Final model partial regression coefficients that meet statistical significance to reject null hypothesis at the $\alpha = 0.05$ level: “forearm”, “wrist”, “thigh”, “abdomen”, “neck”, “weight”, and “intercept” terms.

Final model partial regression coefficients that do not meet statistical significance to reject null hypothesis at the $\alpha = 0.05$ level: “ankle”, “hip”, and “neck”.

The statistical significance test is to determine whether there is evidence to reject the null hypothesis that the estimated beta coefficient is equal to zero (statistical significance referring to being statistically significant from zero).

4.

Summarize the forward selection method of model selection by providing:

```
null.fat <- lm(fat~1, data=bodyfat)
summary(null.fat)
```

```
##
## Call:
## lm(formula = fat ~ 1, data = bodyfat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0602  -6.3602   0.3398   5.8398  21.3398
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.7602     0.7056   26.59   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.137 on 132 degrees of freedom
```

```
for.fat <- step(null.fat, scope=formula(full.fat), direction="forward")
```

```
## Start:  AIC=558.66
## fat ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + abdomen  1    5458.3 3282.3 430.39
## + chest    1    3798.6 4942.0 484.82
## + hip      1    2877.2 5863.4 507.56
## + weight   1    2609.0 6131.6 513.51
## + thigh    1    2501.6 6239.1 515.82
## + knee     1    2190.0 6550.6 522.30
## + biceps   1    1518.4 7222.3 535.28
## + neck     1    1465.1 7275.6 536.26
## + forearm  1    1058.1 7682.6 543.50
## + age      1     881.6 7859.0 546.52
## + wrist    1     712.5 8028.2 549.35
## + ankle    1     417.2 8323.5 554.15
## + height   1     253.6 8487.1 556.74
## <none>                 8740.6 558.66
##
## Step:  AIC=430.39
## fat ~ abdomen
##
##           Df Sum of Sq    RSS    AIC
## + weight   1    801.13 2481.2 395.18
## + hip      1    456.19 2826.1 412.49
## + wrist    1    421.86 2860.5 414.10
## + neck     1    415.43 2866.9 414.39
## + height   1    325.14 2957.2 418.52
## + age      1    231.19 3051.1 422.68
## + knee     1    141.48 3140.8 426.53
## + chest    1    138.56 3143.8 426.66
## + biceps   1    123.33 3159.0 427.30
## + thigh    1    107.32 3175.0 427.97
## + ankle    1     75.56 3206.8 429.29
## + forearm  1     58.59 3223.7 430.00
## <none>                 3282.3 430.39
##
## Step:  AIC=395.18
## fat ~ abdomen + weight
##
##           Df Sum of Sq    RSS    AIC
## + forearm  1    83.494 2397.7 392.62
## + thigh    1    59.525 2421.7 393.95
## + height   1    51.288 2429.9 394.40
## + wrist    1    47.540 2433.7 394.60
## + knee     1    39.778 2441.4 395.03
## <none>                 2481.2 395.18
## + neck     1    35.640 2445.5 395.25
## + biceps   1    25.445 2455.7 395.81
## + ankle    1    25.124 2456.1 395.82
## + chest    1     6.025 2475.2 396.85
## + age      1     5.849 2475.3 396.86
## + hip      1     2.685 2478.5 397.03
##
## Step:  AIC=392.62
## fat ~ abdomen + weight + forearm
##
##          Df Sum of Sq    RSS    AIC
## + wrist   1   123.394 2274.3 387.60
## + neck    1    73.328 2324.4 390.49
## + height  1    53.666 2344.0 391.61
## + thigh   1    42.740 2354.9 392.23
## <none>                2397.7 392.62
## + knee    1    33.551 2364.1 392.75
## + ankle   1    30.348 2367.3 392.93
## + age     1     6.881 2390.8 394.24
## + biceps  1     2.055 2395.6 394.51
## + chest   1     0.285 2397.4 394.61
## + hip     1     0.039 2397.7 394.62
##
## Step:  AIC=387.6
## fat ~ abdomen + weight + forearm + wrist
##
##          Df Sum of Sq    RSS    AIC
## + age     1    88.513 2185.8 384.32
## + ankle   1    57.631 2216.7 386.18
## + knee    1    37.680 2236.6 387.38
## <none>                2274.3 387.60
## + neck    1    29.040 2245.3 387.89
## + height  1    25.196 2249.1 388.12
## + thigh   1     7.719 2266.6 389.15
## + hip     1     7.715 2266.6 389.15
## + biceps  1     0.410 2273.9 389.57
## + chest   1     0.364 2273.9 389.58
##
## Step:  AIC=384.32
## fat ~ abdomen + weight + forearm + wrist + age
##
##          Df Sum of Sq    RSS    AIC
## + ankle   1    66.431 2119.3 382.21
## + thigh   1    42.665 2143.1 383.70
## + neck    1    39.176 2146.6 383.91
## + knee    1    37.367 2148.4 384.02
## <none>                2185.8 384.32
## + height  1    23.983 2161.8 384.85
## + chest   1     1.121 2184.7 386.25
## + hip     1     0.429 2185.3 386.29
## + biceps  1     0.089 2185.7 386.31
##
## Step:  AIC=382.21
## fat ~ abdomen + weight + forearm + wrist + age + ankle
##
##          Df Sum of Sq    RSS    AIC
## + thigh   1    43.457 2075.9 381.46
## <none>                2119.3 382.21
## + neck    1    27.516 2091.8 382.47
## + knee    1    23.248 2096.1 382.75
## + height  1    20.965 2098.4 382.89
## + chest   1     0.815 2118.5 384.16
## + hip     1     0.529 2118.8 384.18
## + biceps  1     0.214 2119.1 384.20
##
## Step:  AIC=381.46
## fat ~ abdomen + weight + forearm + wrist + age + ankle + thigh
##
##          Df Sum of Sq    RSS    AIC
## + neck    1    32.584 2043.3 381.35
## <none>                2075.9 381.46
## + hip     1    19.995 2055.9 382.17
## + knee    1     8.674 2067.2 382.90
## + height  1     4.891 2071.0 383.14
## + biceps  1     1.479 2074.4 383.36
## + chest   1     0.141 2075.8 383.45
##
## Step:  AIC=381.35
## fat ~ abdomen + weight + forearm + wrist + age + ankle + thigh +
##     neck
##
##          Df Sum of Sq    RSS    AIC
## + hip     1    32.303 2011.0 381.23
## <none>                2043.3 381.35
## + knee    1     3.762 2039.5 383.11
## + height  1     3.447 2039.9 383.13
## + chest   1     0.896 2042.4 383.29
## + biceps  1     0.557 2042.8 383.32
```

```
##
## Step:  AIC=381.23
## fat ~ abdomen + weight + forearm + wrist + age + ankle + thigh +
##     neck + hip
##
##          Df Sum of Sq    RSS    AIC
## <none>                2011.0 381.23
## + height  1   11.0257 2000.0 382.50
## + knee    1    3.5666 2007.4 383.00
## + biceps  1    0.7736 2010.2 383.18
## + chest   1    0.2592 2010.8 383.22
```

```
summary(for.fat)
```

```
##
## Call:
## lm(formula = fat ~ abdomen + weight + forearm + wrist + age +
##     ankle + thigh + neck + hip, data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5547 -2.8437 -0.2409  2.6936  8.8349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30.91894   15.50171  -1.995  0.04830 *
## abdomen       0.96082    0.09161  10.488  < 2e-16 ***
## weight       -0.12734    0.05162  -2.467  0.01500 *
## forearm       0.99644    0.32106   3.104  0.00237 **
## wrist        -2.24539    0.73463  -3.056  0.00275 **
## age           0.11847    0.04032   2.938  0.00394 **
## ankle         0.41844    0.23095   1.812  0.07245 .
## thigh         0.38262    0.17438   2.194  0.03010 *
## neck         -0.44046    0.26581  -1.657  0.10006
## hip          -0.24520    0.17445  -1.406  0.16236
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.043 on 123 degrees of freedom
## Multiple R-squared:  0.7699, Adjusted R-squared:  0.7531
## F-statistic: 45.73 on 9 and 123 DF,  p-value: < 2.2e-16
```

(a)

An ordered list of which variable was added to the model at each step;

Step 1: “abdomen” added Step 2: “weight” added Step 3: “forearm” added Step 4: “wrist” added Step 5: “age” added Step 6: “ankle” added Step 7: “thigh” added Step 8: “neck” added Step 9: “hip” added

(b)

A list of which variables never entered the final model;

Never entered the final model: “height”, “chest”, “knee”, and “biceps”.

(c)

A summary of the partial regression coefficients effects tests for the final model.

Final model partial regression coefficients that meet statistical significance to reject null hypothesis at the $\alpha = 0.05$ level: “abdomen”, “weight”, “forearm”, “wrist”, “age”, “thigh”, and “intercept”

Final model partial regression coefficients that do not meet statistical significance to reject null hypothesis at the $\alpha = 0.05$ level: “ankle”, “hip”, and “neck”.

The statistical significance test is to determine whether there is evidence to reject the null hypothesis that the estimated beta coefficient is equal to zero (statistical significance referring to being statistically significant from zero).

5.

Summarize the all-possible-subsets method of model selection by providing:

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.4.2
```

```
all.subsets <- regsubsets(fat~., data=bodyfat, method="exhaustive")
summary(all.subsets)
```

```
## Subset selection object
## Call: regsubsets.formula(fat ~ ., data = bodyfat, method = "exhaustive")
## 13 Variables  (and intercept)
##         Forced in Forced out
## age         FALSE      FALSE
## weight      FALSE      FALSE
## height      FALSE      FALSE
## neck        FALSE      FALSE
## chest       FALSE      FALSE
## abdomen     FALSE      FALSE
## hip         FALSE      FALSE
## thigh       FALSE      FALSE
## knee        FALSE      FALSE
## ankle       FALSE      FALSE
## biceps      FALSE      FALSE
## forearm     FALSE      FALSE
## wrist       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          age weight height neck chest abdomen hip thigh knee ankle biceps
## 1  ( 1 ) " " " "    " "    " "  " "   "*"     " " " "   " "  " "   " "   
## 2  ( 1 ) " " "*"    " "    " "  " "   "*"     " " " "   " "  " "   " "   
## 3  ( 1 ) " " "*"    " "    " "  " "   "*"     " " " "   " "  " "   " "   
## 4  ( 1 ) " " "*"    " "    " "  " "   "*"     " " " "   " "  " "   " "   
## 5  ( 1 ) "*" "*"    " "    " "  " "   "*"     " " " "   " "  " "   " "   
## 6  ( 1 ) "*" "*"    " "    " "  " "   "*"     " " " "   " "  "*"   " "   
## 7  ( 1 ) "*" "*"    " "    " "  " "   "*"     " " "*"   " "  "*"   " "   
## 8  ( 1 ) "*" "*"    " "    "*"  " "   "*"     " " "*"   " "  "*"   " "   
##          forearm wrist
## 1  ( 1 ) " "     " "  
## 2  ( 1 ) " "     " "  
## 3  ( 1 ) "*"     " "  
## 4  ( 1 ) "*"     "*"  
## 5  ( 1 ) "*"     "*"  
## 6  ( 1 ) "*"     "*"  
## 7  ( 1 ) "*"     "*"  
## 8  ( 1 ) "*"     "*"
```

```
summary(all.subsets)$adjr2
```

```
## [1] 0.6216099 0.7117651 0.7193053 0.7316709 0.7400833 0.7459827 0.7492008
## [8] 0.7511467
```

(a)

Which model would you choose based on the adjusted R^2 values?

The highest adjusted R^2 value is 0.7511467 for the model 8, the model that uses: "ankle", "forearm", "wrist", "thigh", "abdomen", "neck", "weight", and "age", including an intercept term

Note: Of the models considered using the above method, I stand by my answer. However, it is worth noting that we can consider models with more explanatory variables (13 possible explanatory variables to chose from, and from playing around I found a "better"/larger Adjusted R squared value for a model with 9 explanatory variables). The above code only looks at up to 8 explanatory variables in its output, without messing around with the `nvmax` parameter in the function.

(b)

Which model would you choose based on the Mallow's C_p criteria?

```r
summary_object <- summary(all.subsets)

summary(all.subsets)$cp
```

```
## [1] 66.743373 20.967366 17.988124 12.629401  9.350862  7.389186  6.797574
## [8]  6.854376
```

```r
included_matrix <- summary_object$which
num_vars_per_model <- rowSums(included_matrix) + 1
num_vars_per_model
```

```
##  1  2  3  4  5  6  7  8
##  3  4  5  6  7  8  9 10
```

```r
abs(summary(all.subsets)$cp - num_vars_per_model)
```

```
##          1          2          3          4          5          6          7
## 63.7433734 16.9673662 12.9881241  6.6294005  2.3508623  0.6108144  2.2024261
##          8
##  3.1456237
```

The "best" Mallow's C_p value is the value that corresponds the closest to the number of explanatory variables in the model (including an intercept, also more "magnitude" in terms of signage). Given the above output, we'd choose model 6, which has the explanatory variables: "ankle", "forearm", "wrist", "abdomen", "weight", and "age", including an intercept term.

(c)

Which model would you choose based on the BIC values?

```r
summary(all.subsets)$bic
```

```
## [1] -120.4841 -152.8088 -152.4711 -154.6078 -154.9971 -154.2116 -152.0768
## [8] -149.2907
```

```
min(summary(all.subsets)$bic)
```

```
## [1] -154.9971
```

```
which.min(summary(all.subsets)$bic)
```

```
## [1] 5
```

The “best” BIC value is the one that is minimized across the models considered. The BIC value is minimized in model 5, the model that uses the explanatory variables: “forearm”, “wrist”, “abdomen”, “weight”, and “age”, including an intercept term.

6.

Interpret the values of the estimated regression coefficients in the context of the study for:

```r
ageCat <- vector(mode="character", length=length(bodyfat$age))
ageCat[bodyfat$age<39] = "under39" # lower quartile
ageCat[bodyfat$age>52] = "over52" # upper quartile
ageCat[bodyfat$age>38 & bodyfat$age<53] = "mid" # middle 50%
bodyfat = cbind(bodyfat, ageCat)
```

```r
cat.fat <- lm(fat ~ ageCat + weight + neck + abdomen + hip + thigh +
ankle + forearm + wrist, data = bodyfat)
summary(cat.fat)
```

```
##
## Call:
## lm(formula = fat ~ ageCat + weight + neck + abdomen + hip + thigh +
##     ankle + forearm + wrist, data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9891 -2.8403 -0.1501  2.5172  8.7621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -30.14180   15.64789  -1.926  0.05640 .
## ageCatover52    1.98494    0.97316   2.040  0.04354 *
## ageCatunder39  -2.06578    0.96722  -2.136  0.03469 *
## weight         -0.14417    0.05102  -2.825  0.00552 **
## neck           -0.34301    0.26643  -1.287  0.20038
## abdomen         0.98274    0.08612  11.412  < 2e-16 ***
## hip            -0.22144    0.17472  -1.267  0.20743
## thigh           0.35881    0.17202   2.086  0.03908 *
## ankle           0.48082    0.23253   2.068  0.04078 *
## forearm         0.99011    0.31998   3.094  0.00245 **
## wrist          -2.26634    0.73520  -3.083  0.00254 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.04 on 122 degrees of freedom
## Multiple R-squared:  0.7722, Adjusted R-squared:  0.7535
## F-statistic: 41.35 on 10 and 122 DF,  p-value: < 2.2e-16
```

(a)

The two values corresponding to the categorical age variable;

Note: - “Baseline” for comparisons is male individuals Age between 39 to 52 - “Individuals” in the context of the study are 133 men, such that I’d argue the “population” we interpret with regards to is males - “Fat” is not noted to have a particular unit in the Lab notes, though “weight” is denoted in “lbs”. For that reason I refer to units of “fat” and not “pounds of”

We reject the null hypothesis at the $\alpha = 0.05$ level that there is no difference in mean bodyfat between male individuals under 39 years of age compared to mean bodyfat for individuals between 39 to 52 years of age (conditional mean of bodyfat, units were not provided for the response). This is to say we have evidence in favor of the alternative hypothesis, specifically that all else being equal, we expect individuals under 39 years of age to have 2.06578 units less bodyfat than individuals between 39 to 52 years of age.

We reject the null hypothesis at the $\alpha = 0.05$ level that there is no difference in mean bodyfat between male individuals over 52 years of age compared to mean bodyfat for individuals between 39 to 52 years of age (again, conditional mean with units of response variable not specified). This is to say we have evidence in favor of the alternative hypothesis, specifically that all else being equal, we expect individuals over 52 years of age to have 1.98494 units more bodyfat than individuals between 39 to 52 years of age.

(b)

One of the values corresponding to the quantitative variable of your choice.

We have evidence at the $\alpha = 0.05$ level to reject the null hypothesis that increasing the circumference of the abdomen is not associated with a change in the bodyfat of a male individual. This is evidence in favor of the alternative hypothesis, specifically that increasing the circumference of the abdomen by 1 cm is associated with an increase of bodyfat of 0.98274 units, all else being equal.

7.

Summarize your findings from examining all the residual plots used to diagnose the MLR model assumptions. Are there any assumptions that aren't met for this analysis?

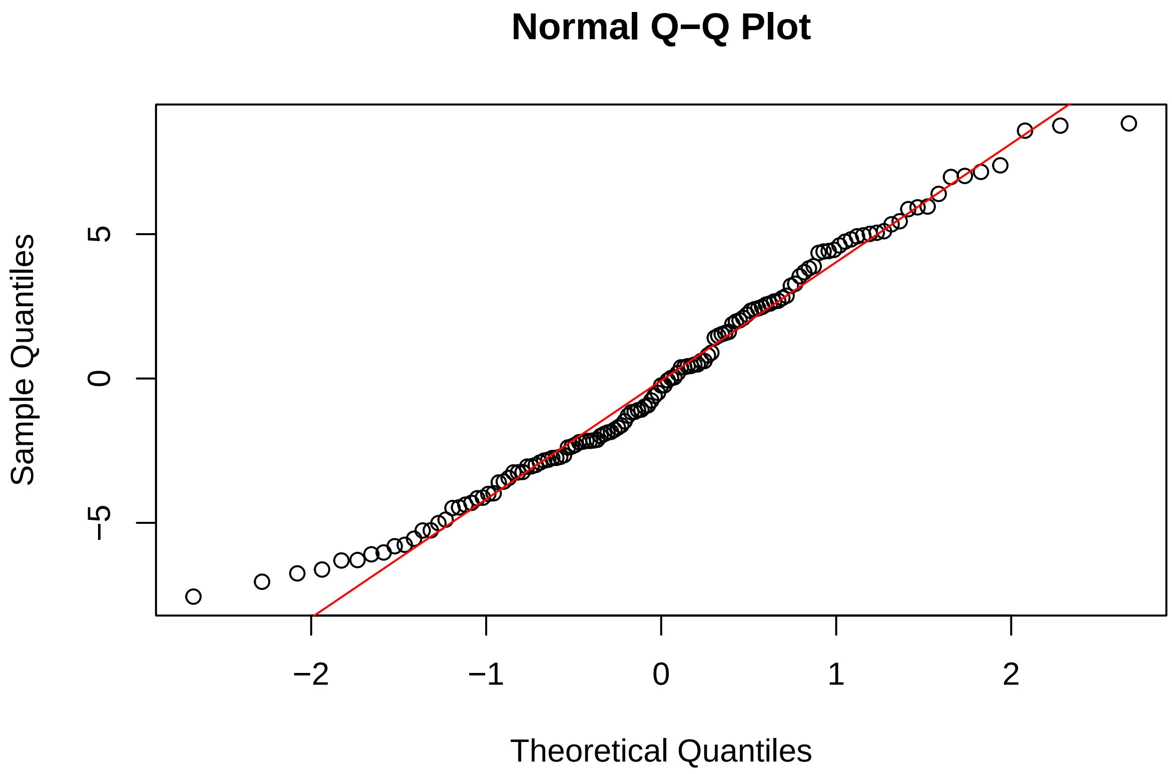
```
best.fat <- back.fat
summary(best.fat)
```

```
##
## Call:
## lm(formula = fat ~ age + weight + neck + abdomen + hip + thigh +
##     ankle + forearm + wrist, data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5547 -2.8437 -0.2409  2.6936  8.8349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30.91894   15.50171  -1.995  0.04830 *
## age           0.11847    0.04032   2.938  0.00394 **
## weight       -0.12734    0.05162  -2.467  0.01500 *
## neck         -0.44046    0.26581  -1.657  0.10006
## abdomen       0.96082    0.09161  10.488  < 2e-16 ***
## hip          -0.24520    0.17445  -1.406  0.16236
## thigh         0.38262    0.17438   2.194  0.03010 *
## ankle         0.41844    0.23095   1.812  0.07245 .
## forearm       0.99644    0.32106   3.104  0.00237 **
## wrist        -2.24539    0.73463  -3.056  0.00275 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.043 on 123 degrees of freedom
## Multiple R-squared:  0.7699, Adjusted R-squared:  0.7531
## F-statistic: 45.73 on 9 and 123 DF,  p-value: < 2.2e-16
```

```
anova(best.fat)
```

```
## Analysis of Variance Table
##
## Response: fat
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## age         1  881.60  881.60  53.9215 2.496e-11 ***
## weight      1 2744.51 2744.51 167.8632 < 2.2e-16 ***
## neck        1  131.48  131.48   8.0417  0.005347 **
## abdomen     1 2551.89 2551.89 156.0821 < 2.2e-16 ***
## hip         1    5.62    5.62   0.3439  0.558683
## thigh       1  170.61  170.61  10.4352  0.001586 **
## ankle       1   21.93   21.93   1.3411  0.249091
## forearm     1   69.26   69.26   4.2362  0.041684 *
## wrist       1  152.74  152.74   9.3421  0.002747 **
## Residuals 123 2011.01   16.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

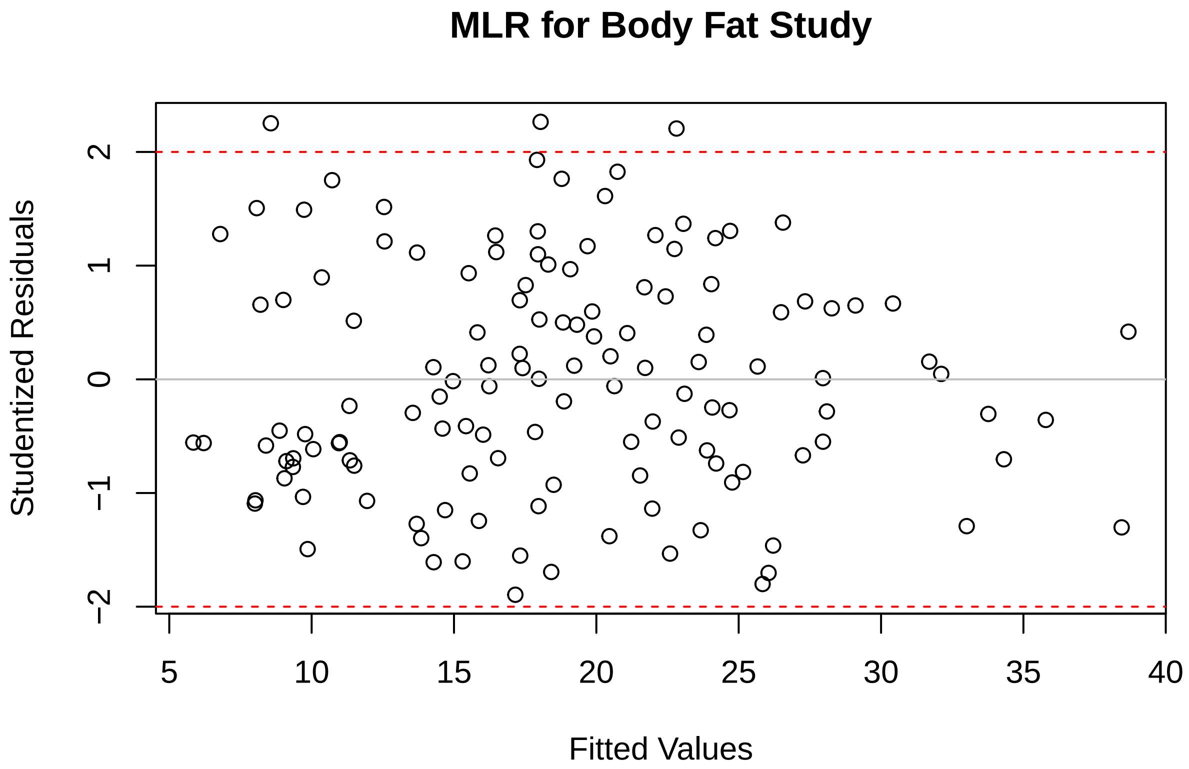
```
qqnorm(best.fat$residuals)
qqline(best.fat$residuals, col="red")
```



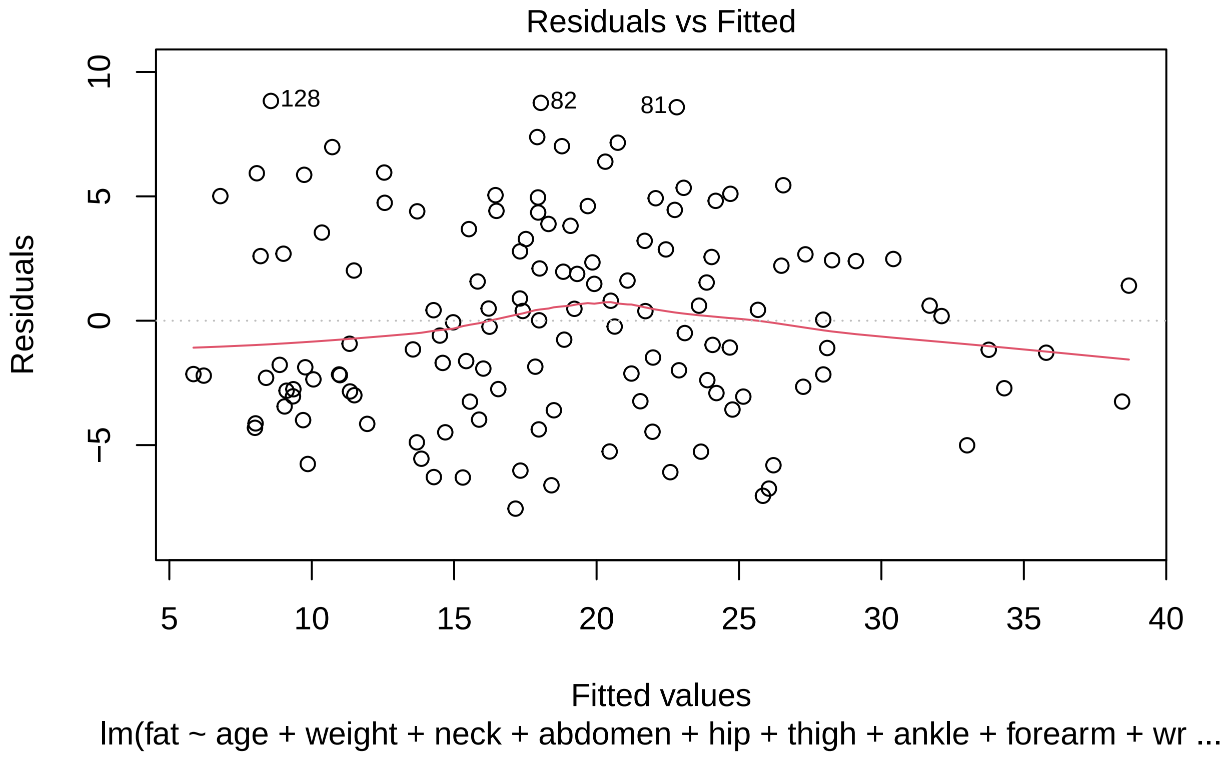
```
library(MASS)
stdresids <- stdres(best.fat)
stdresids[which(abs(stdresids)>2)]
```

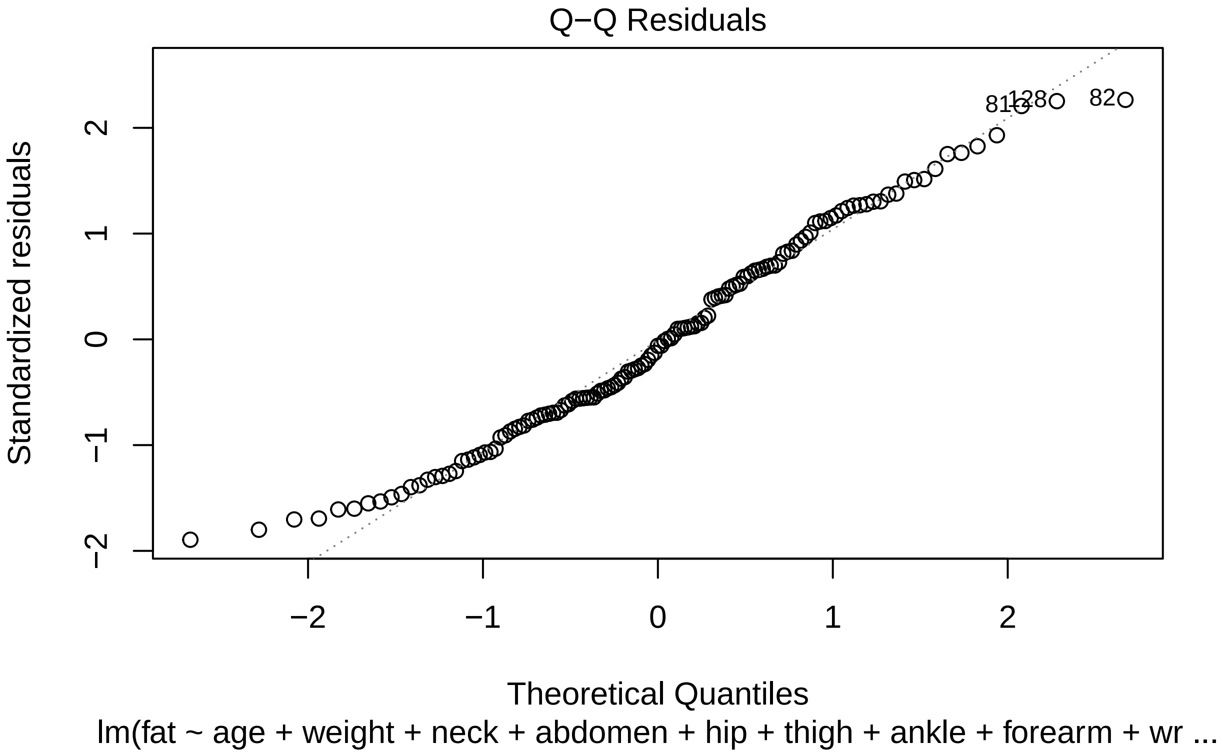
```
##       81       82      128
## 2.206322 2.264770 2.252444
```

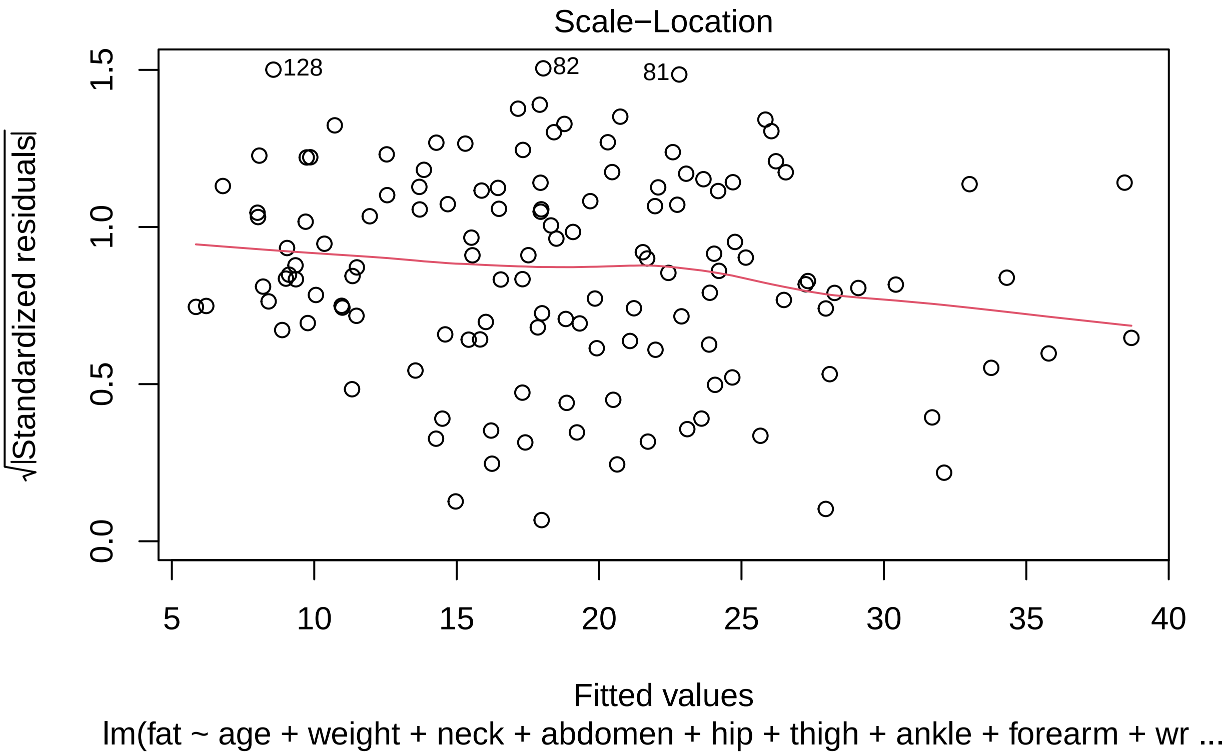
```
plot(best.fat$fitted.values, stdresids, main="MLR for Body Fat Study",
xlab="Fitted Values", ylab="Studentized Residuals")
abline(h=0, col="gray")
abline(h=-2, col="red", lty=2)
abline(h=2, col="red", lty=2)
```

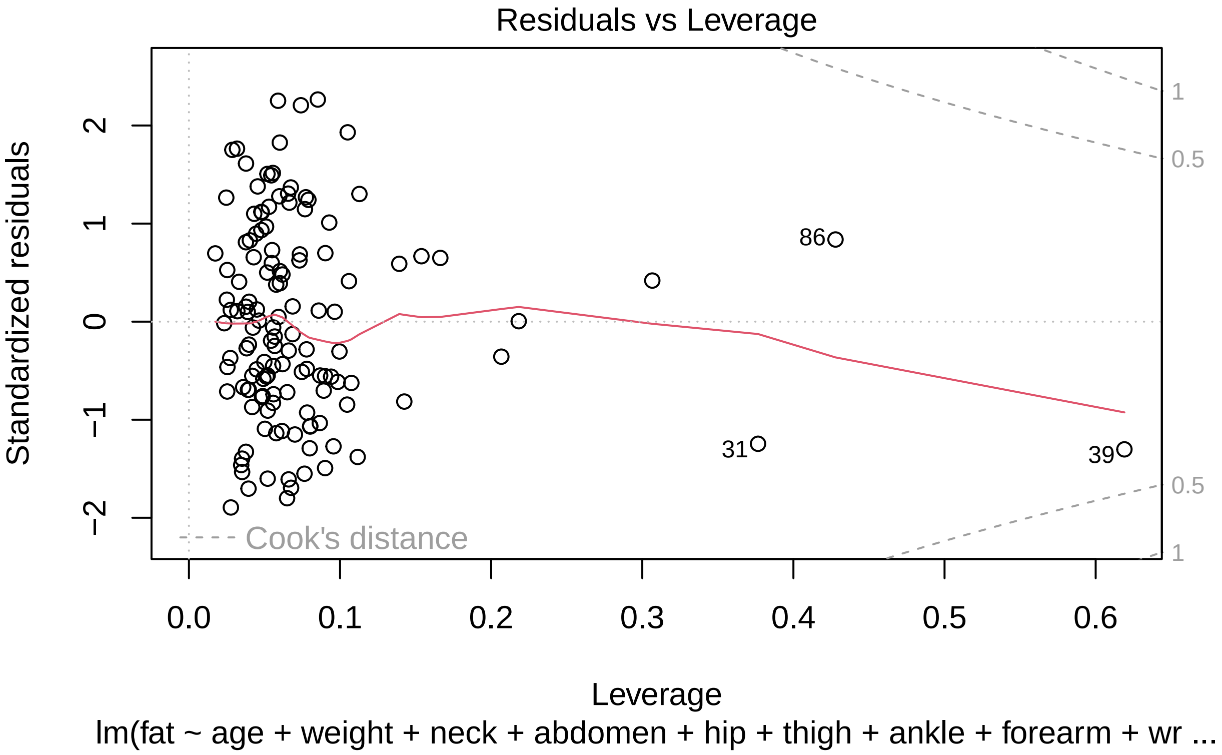


```r
plot(best.fat)
```


Residuals vs Fitted
Residuals
Fitted values
lm(fat ~ age + weight + neck + abdomen + hip + thigh + ankle + forearm + wr ...
128
82
81
10
5
0
−5
5
10
15
20
25
30
35
40


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
Theoretical Quantiles
−2
−1
0
1
2
81
128
82
lm(fat ~ age + weight + neck + abdomen + hip + thigh + ankle + forearm + wr ...


Scale−Location
1.5
1.0
0.5
0.0
√|Standardized residuals|
128
82
81
5
10
15
20
25
30
35
40
Fitted values
lm(fat ~ age + weight + neck + abdomen + hip + thigh + ankle + forearm + wr ...



Overall, the assumptions we are diagnosing are the equal variance, linearity, and normality assumptions. To that end:

Residual Plot: Constant variance and form of the model (linearity) assumptions appear to be met, as the overall spread and distribution of residuals across fitted values appears as a random spread. Specifically, we have a random spread and not a funnel shape (for assessing constant variance), or other types of trend that would indicate a deviation from linearity. Also, we tend to see the same number of positive residuals as we do negative residual values (for assessing form of the model). Overall, our assumptions of equal variance as well as form of the model do not appear to be violated.

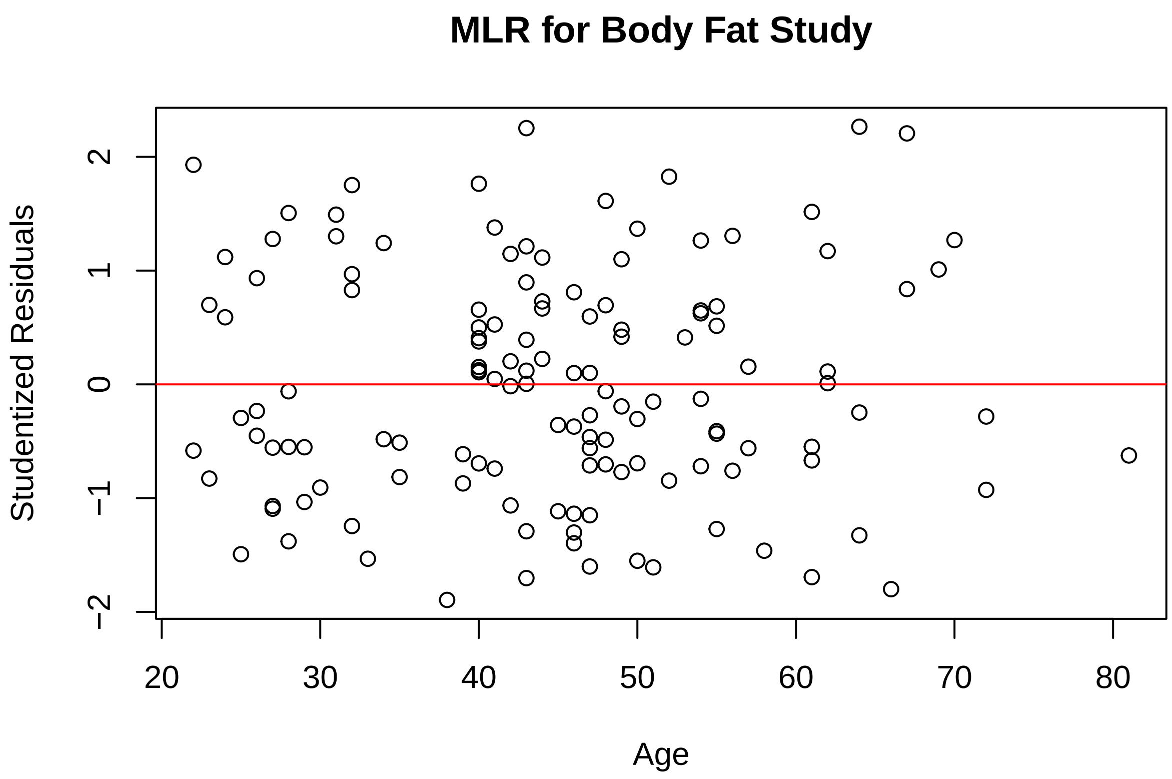
QQ Plot: Residuals track and align well against the reference line, with some slight deviations at the tails of the distribution. This is evidence in favor of the normality assumption not being violated.

Generally though, we do observe some clustering of points together though, as evidenced in the residual plot, such that we have some potential issues with the "fixed X" assumption possibly being violated when looking at residual plots of explanatory variables.

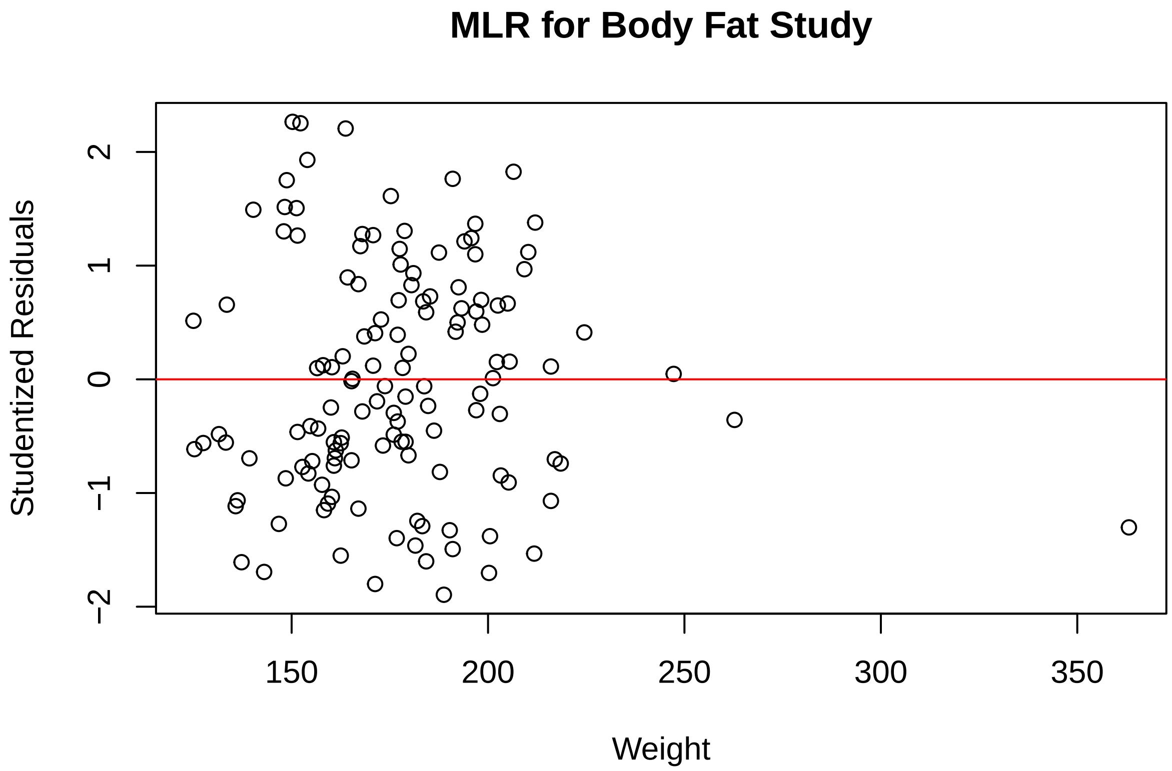
Note: The above findings are consistent when also considering the above plots using the studentized residuals. Furthermore, when we look at the residual plots of each explanatory variable, we also see that there are possibly outliers, which we will further investigate to identify if it is in fact an outlier and possibly a leverage/influential point.

Additional plots given below, maybe they're above, who knows with this knitting business:

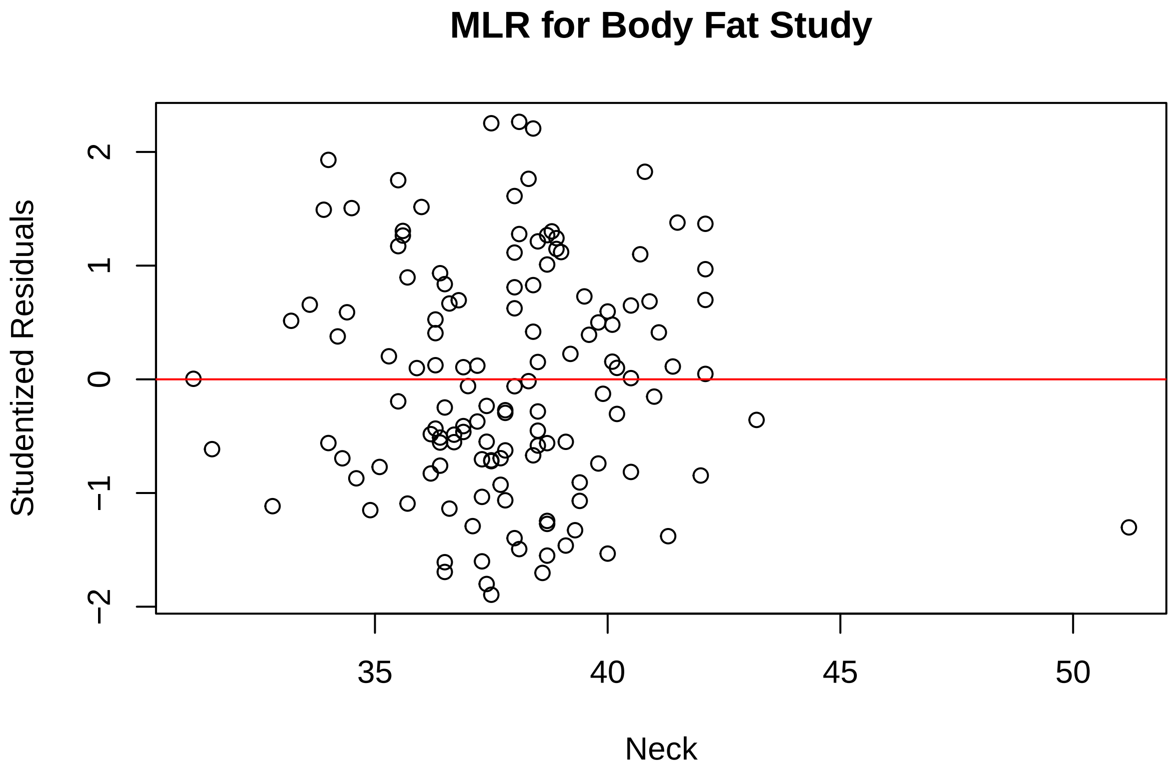
```r
library(MASS)
stdresids <- stdres(best.fat)
plot(best.fat$model$age,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Age",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



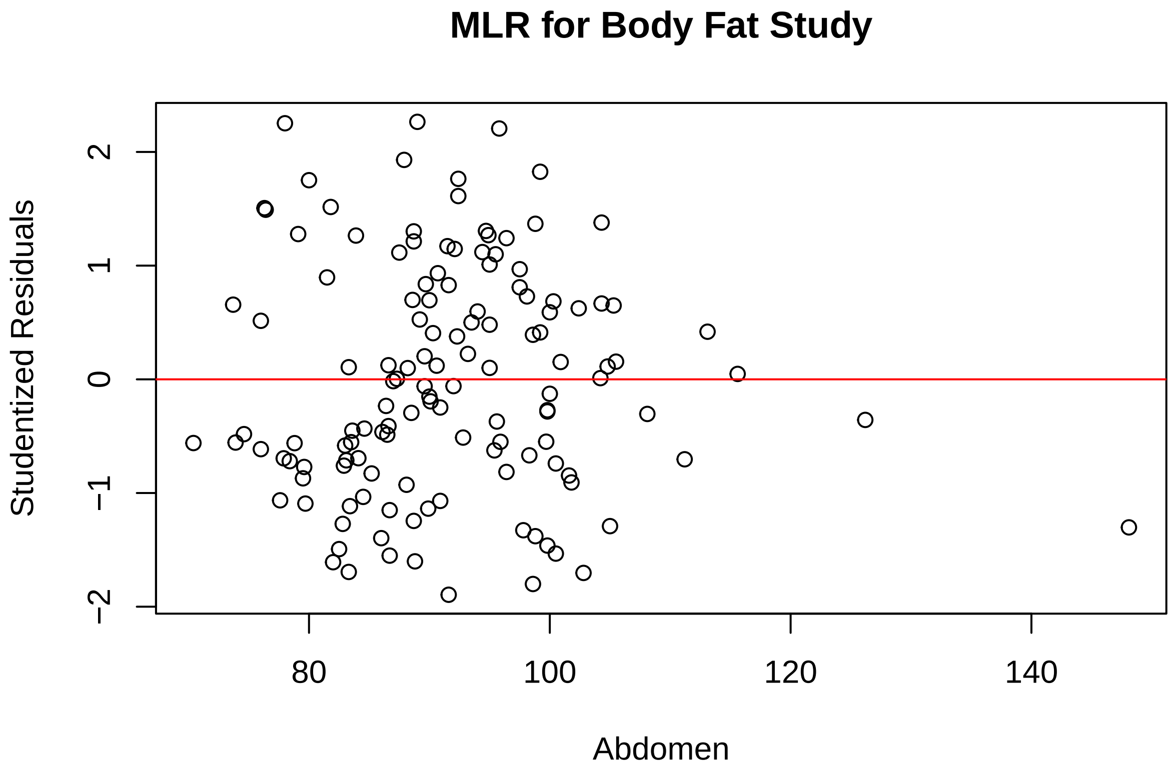
```
plot(best.fat$model$weight,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Weight",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



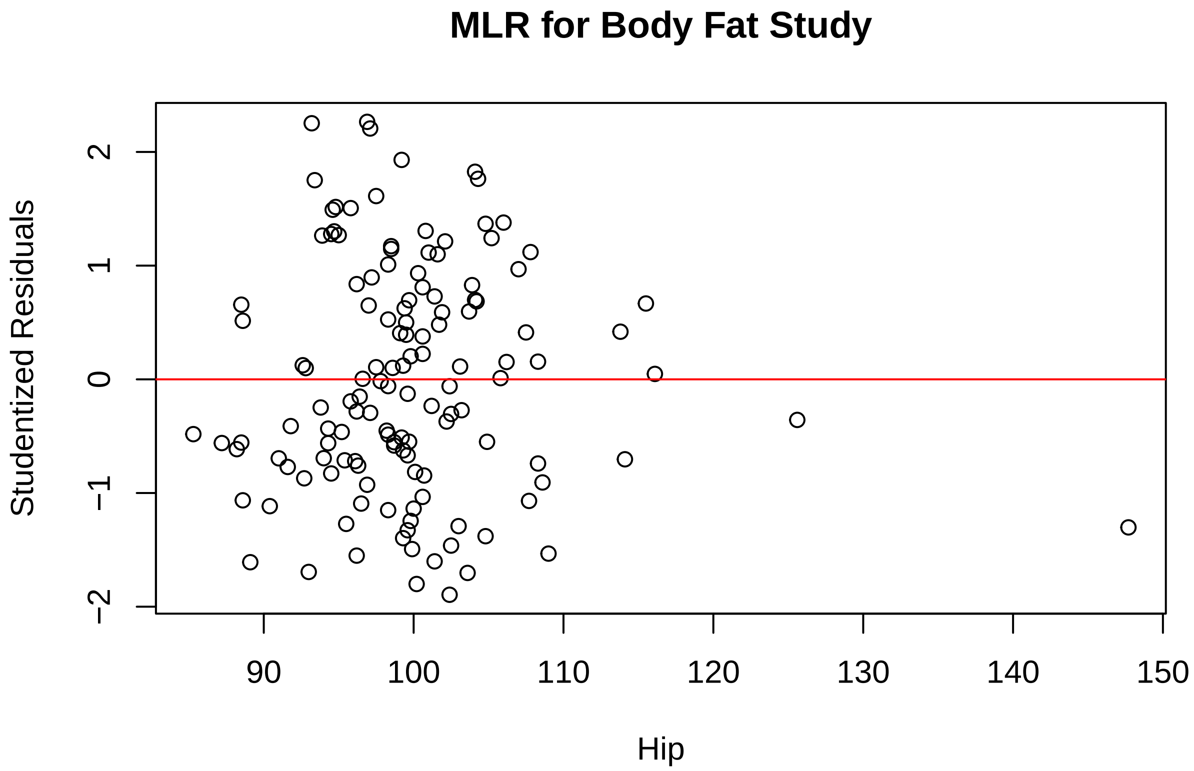
```r
plot(best.fat$model$neck,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Neck",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



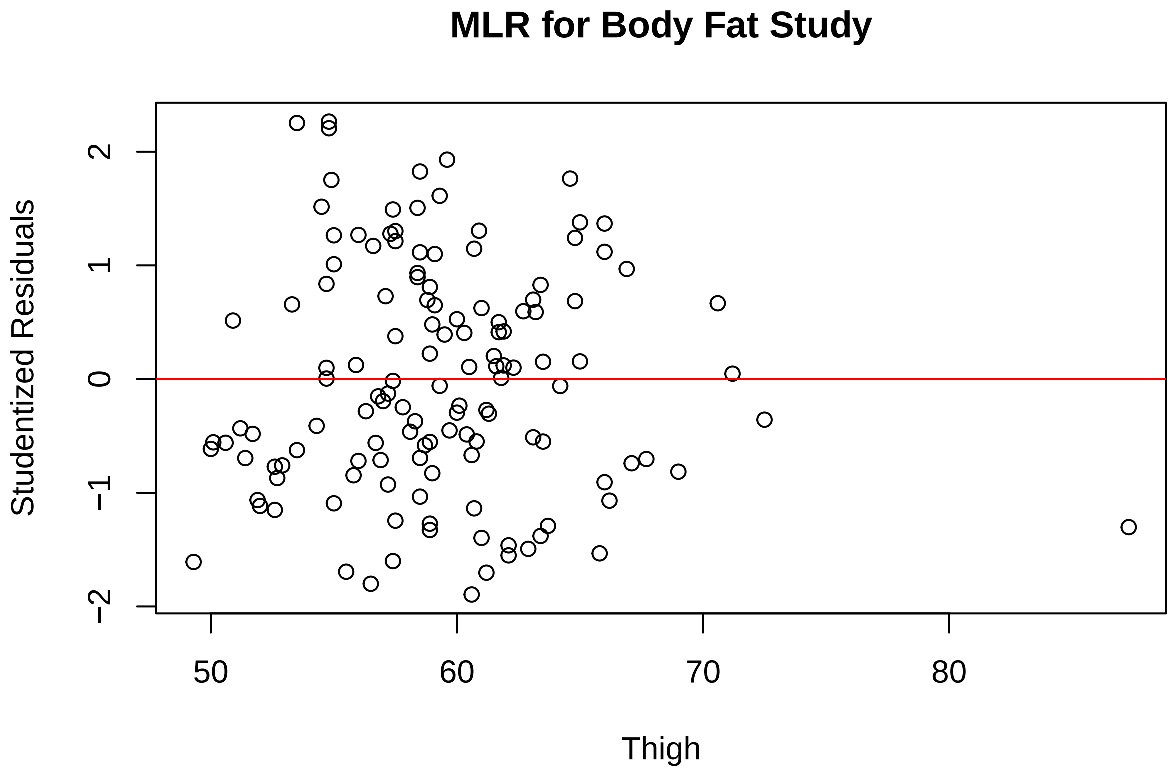
```
plot(best.fat$model$abdomen,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Abdomen",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



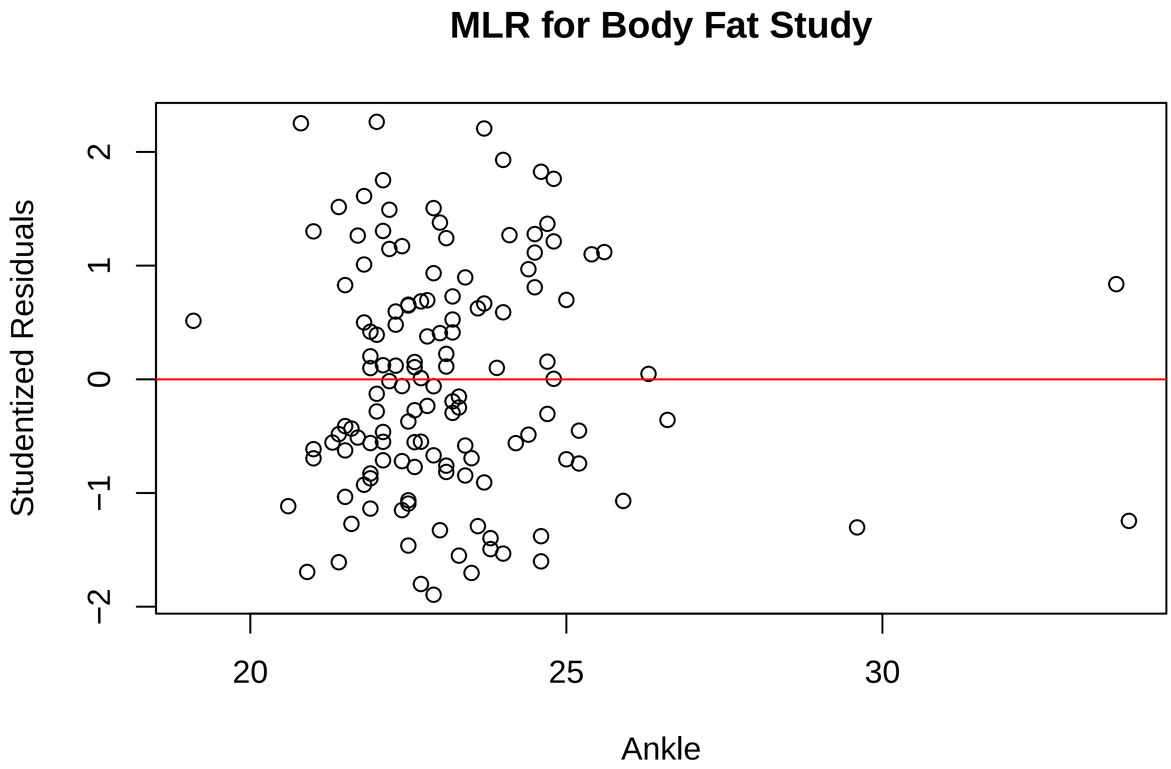
```
plot(best.fat$model$hip,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Hip",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



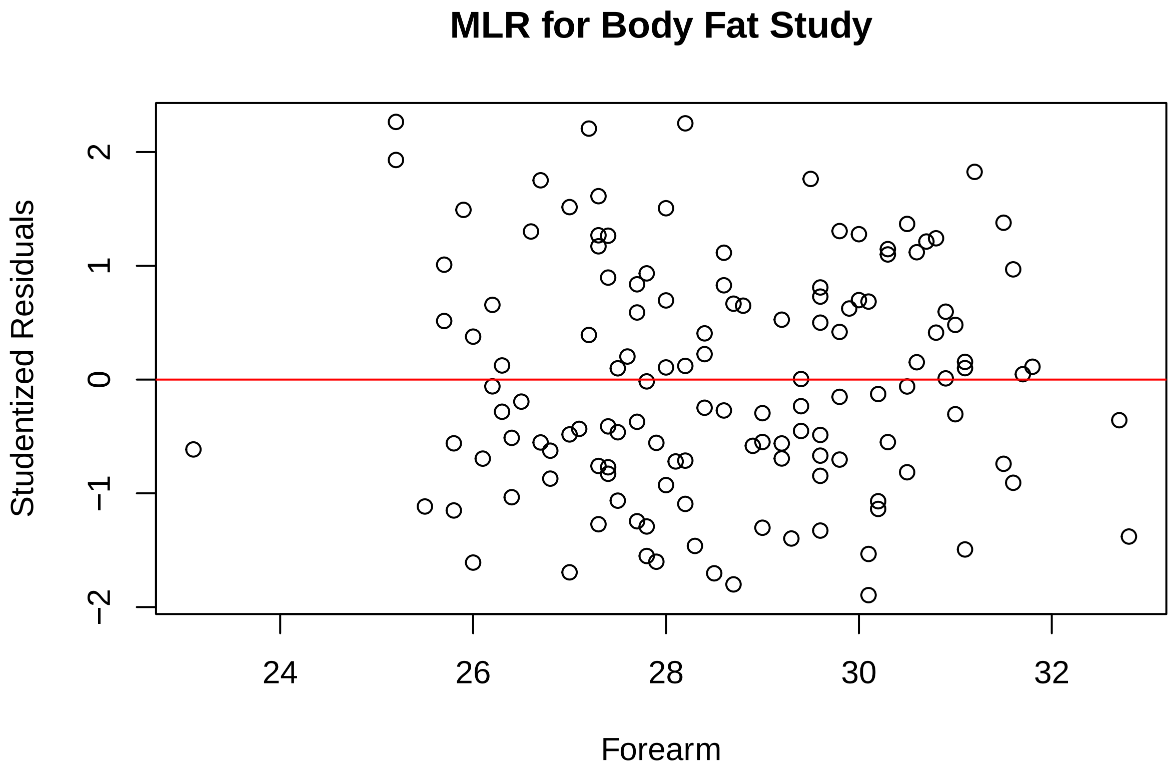
```r
plot(best.fat$model$thigh,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Thigh",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



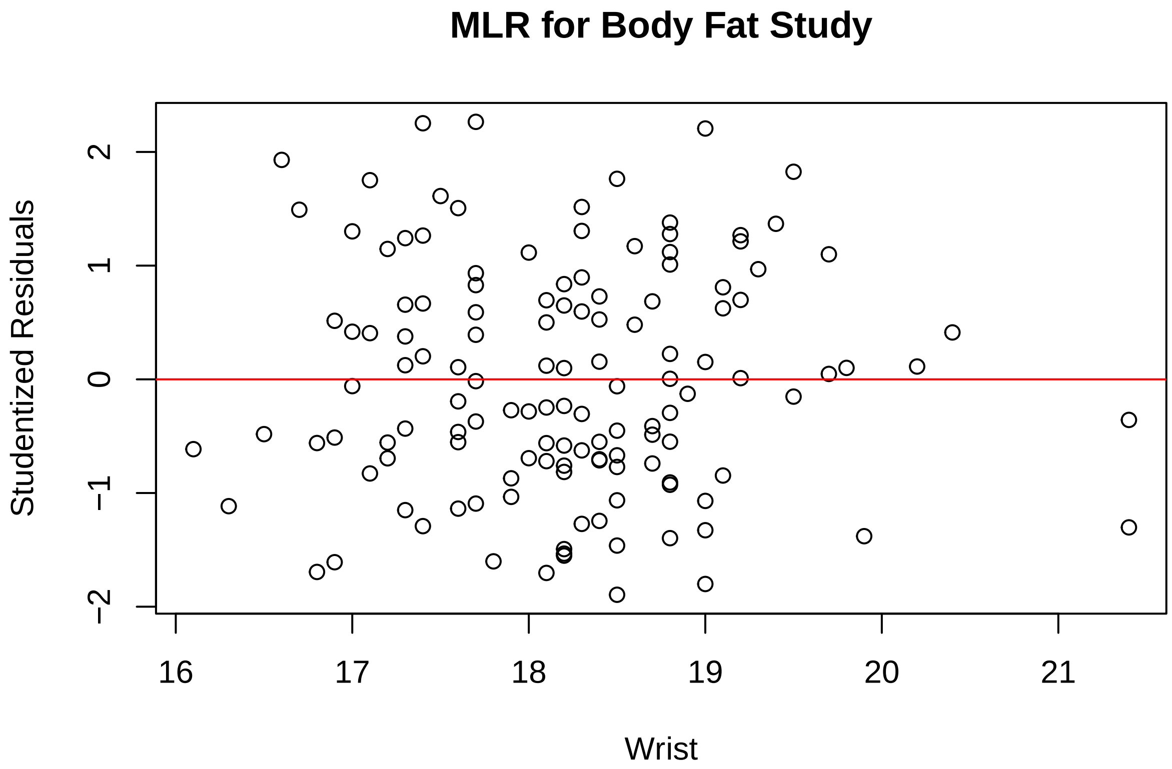
```
plot(best.fat$model$ankle,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Ankle",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



```
plot(best.fat$model$forearm,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Forearm",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```



```
plot(best.fat$model$wrist,
     stdresids,
     main="MLR for Body Fat Study",
     xlab="Wrist",
     ylab="Studentized Residuals")
abline(h=0, col="red")
```


MLR for Body Fat Study
Studentized Residuals
−2
−1
0
1
2
16
17
18
19
20
21
Wrist

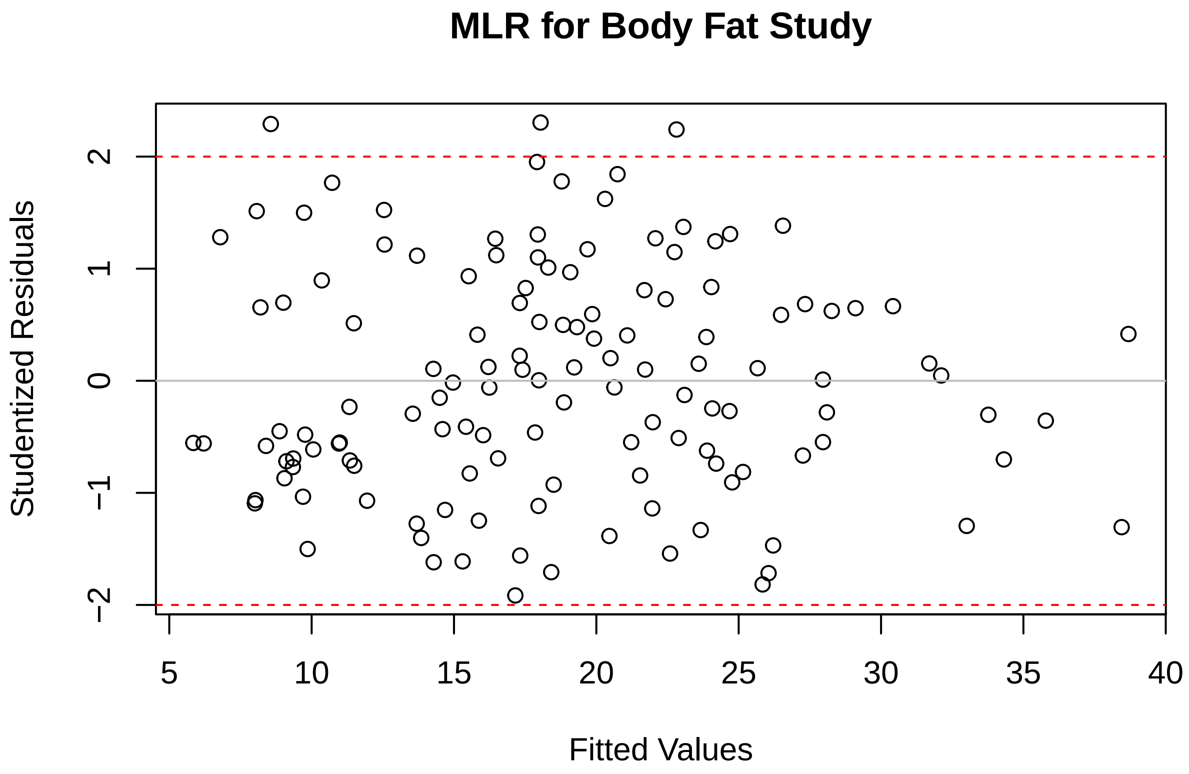
8.

Summarize your findings from examining the case diagnostic values/plots. Are there any outliers, leverage points, or influential observations?

```
stdresids <- studres(best.fat)
stdresids[which(abs(stdresids)>2)]
```

```
##       81       82      128
## 2.242151 2.304098 2.291016
```

```
plot(best.fat$fitted.values, stdresids, main="MLR for Body Fat Study",
xlab="Fitted Values", ylab="Studentized Residuals")
abline(h=0, col="gray")
abline(h=-2, col="red", lty=2)
abline(h=2, col="red", lty=2)
```

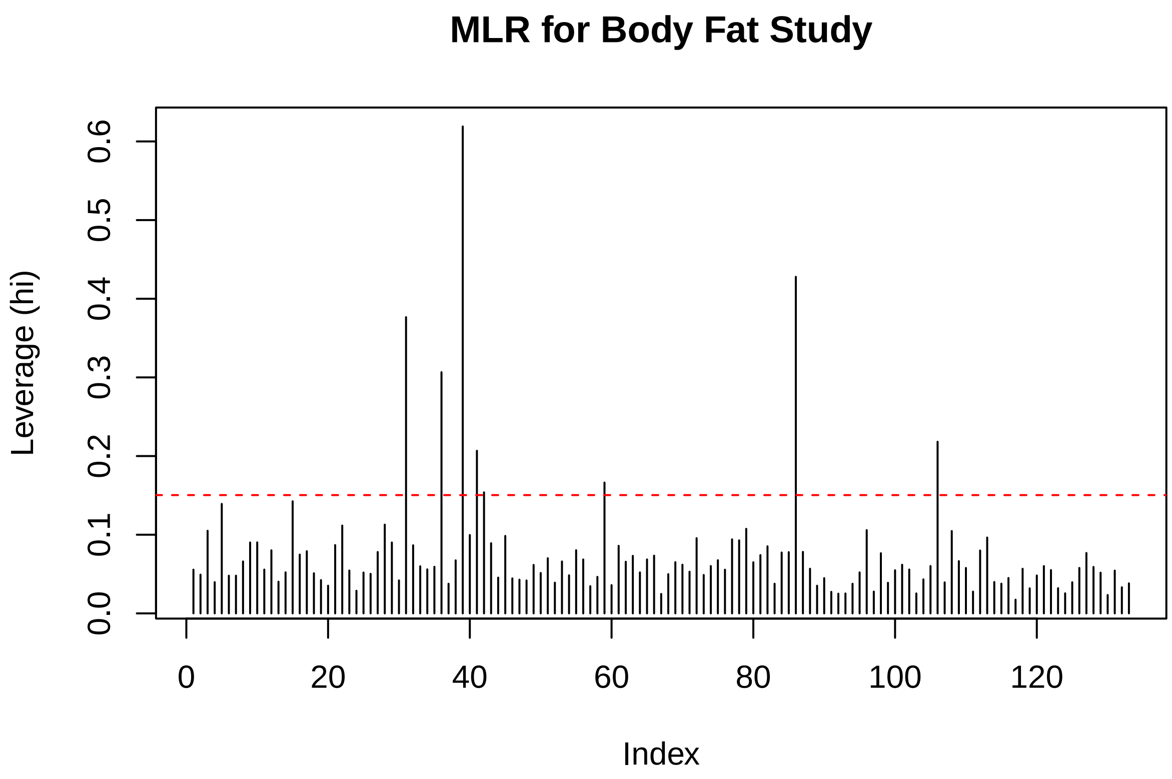


Using the above plot of studentized residuals, we are able to potentially identify outliers in our dataset. We do observe some potential outliers using this method, where we look for residuals with a magnitude greater than 2.

```
leverage <- hatvalues(best.fat)
leverage[which(abs(leverage)>(20/length(leverage)))]
```

```
##        31        36        39        41        42        59        86       106
## 0.3765945 0.3066175 0.6190142 0.2067231 0.1538633 0.1663539 0.4278538 0.2182216
```

```
plot(leverage, type = 'h', main="MLR for Body Fat Study",
ylab="Leverage (hi)")
abline(h=(20/length(leverage)), col="red", lty=2)
```



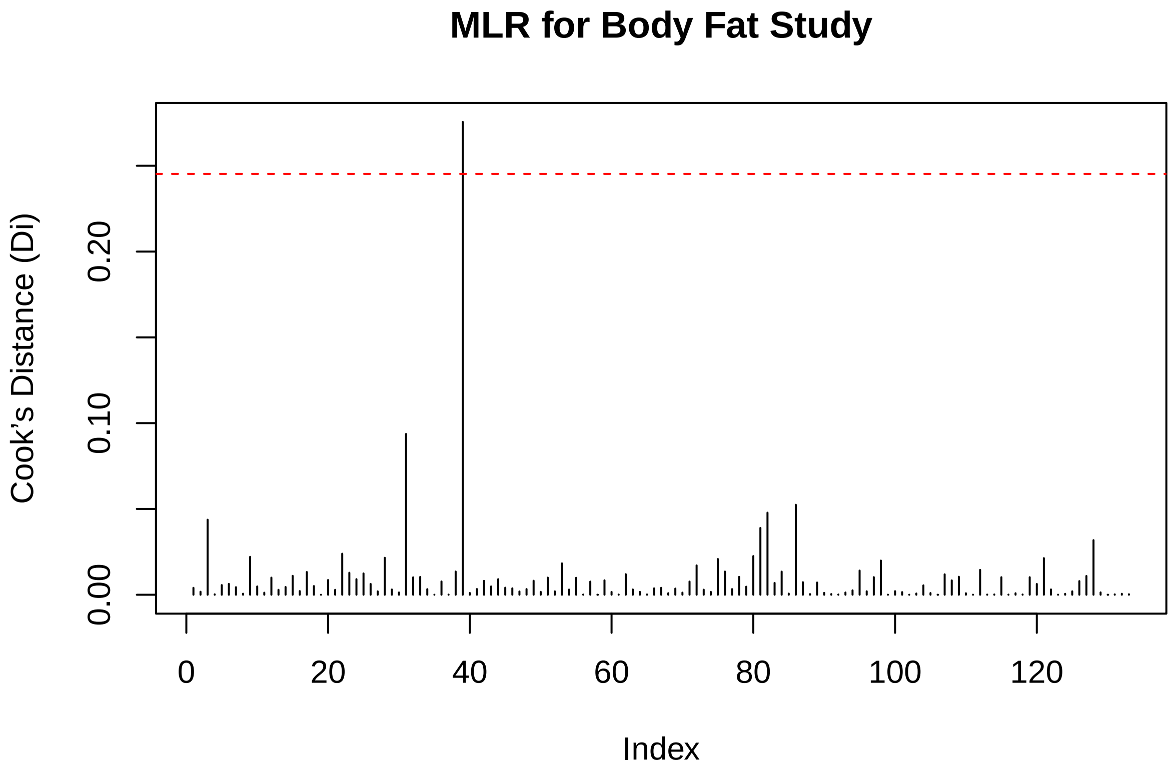
```
20/length(leverage)
```

```
## [1] 0.1503759
```

We can find which observations have leverage exceeding $\pm 2(k+1)/n$ (0.1503759) where k (9) is the number of explanatory variables and n (133) is the total number of observations:

Using the above threshold values for leverage, and the above plot, we see there are a number of observations in our dataset with high leverage (there are leverage points).

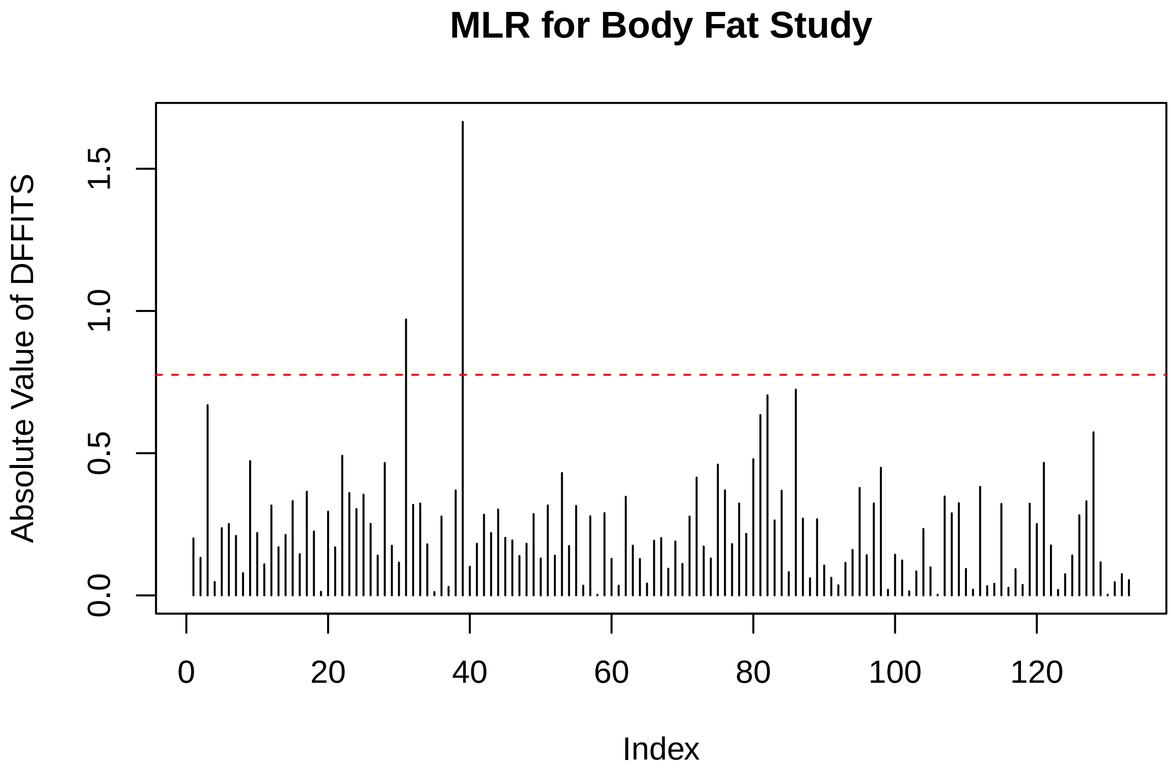
```
cooks <- cooks.distance(best.fat)
# cooks[which(abs(cooks)>(2*sqrt(2/length(cooks))))]
plot(cooks, type = 'h', main="MLR for Body Fat Study",
ylab="Cook's Distance (Di)")
abline(h=2*sqrt(2/length(leverage)), col="red", lty=2)
```



Cook's Threshold value: $\pm 2\sqrt{2/n} \approx 0.2452557$

Cook's Distance is used to evaluate potential influence points. From the above plot and using the specified threshold given, we see there is at least one influence point via this method.

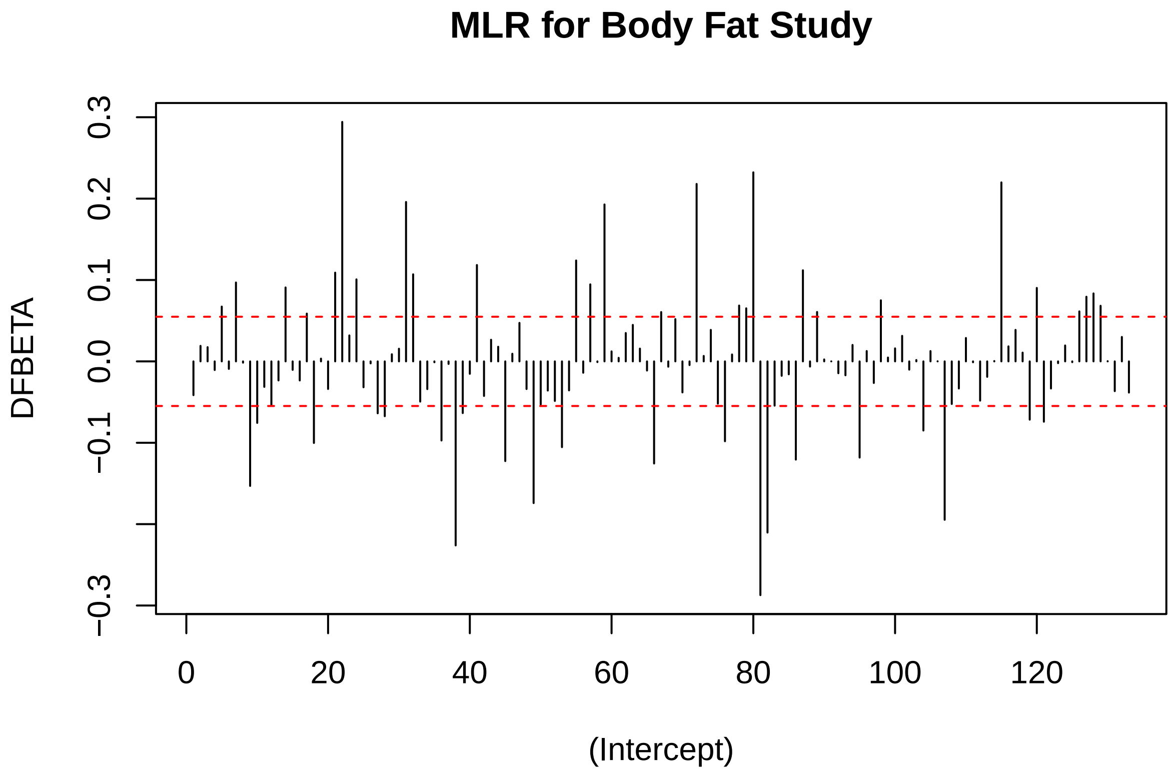
```
dff <- dffits(best.fat)
# dff[which(abs(dff) > 2*sqrt(20/length(dff)))]
plot(abs(dff), type = 'h', main="MLR for Body Fat Study",
ylab="Absolute Value of DFFITS")
abline(h=2*sqrt(20/length(dff)), col="red", lty=2)
```



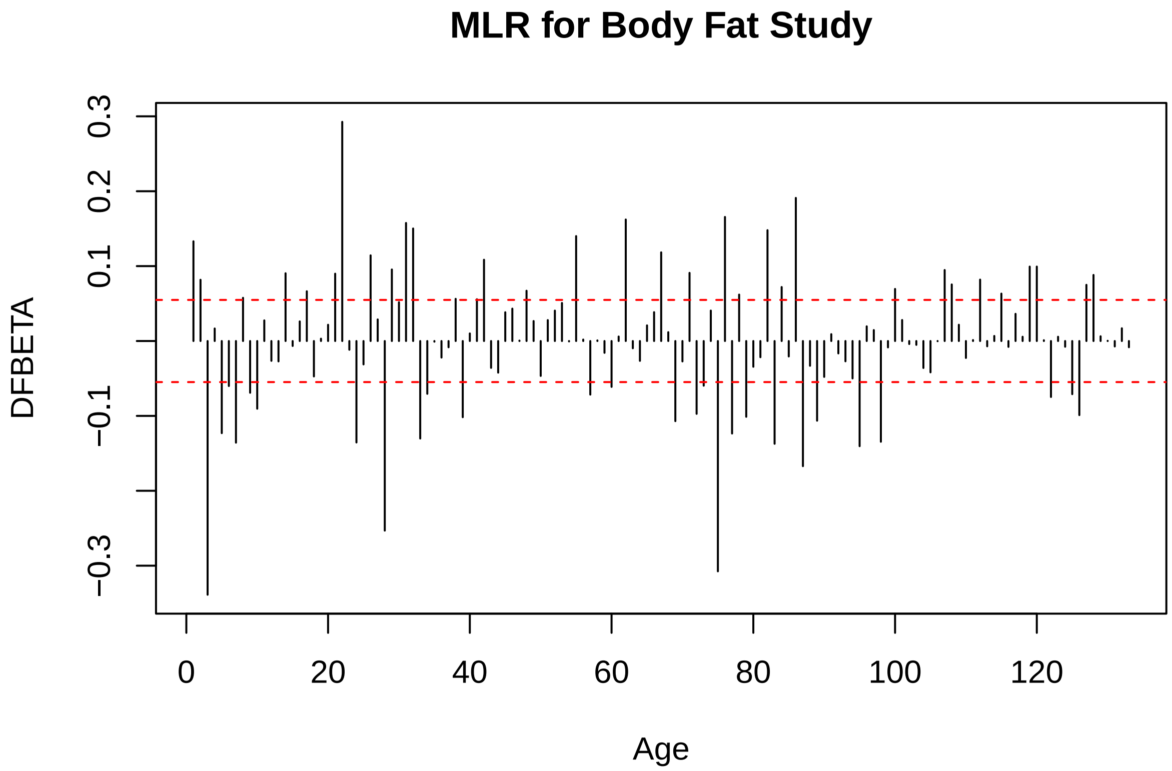
Another way to check for potential influence points is using DFFITS, with the threshold value of $\pm 2\sqrt{(k+1)/n} \approx 0.1503759$

Using the above method, we see there are now potentially two influential points in the dataset used in this study/lab.

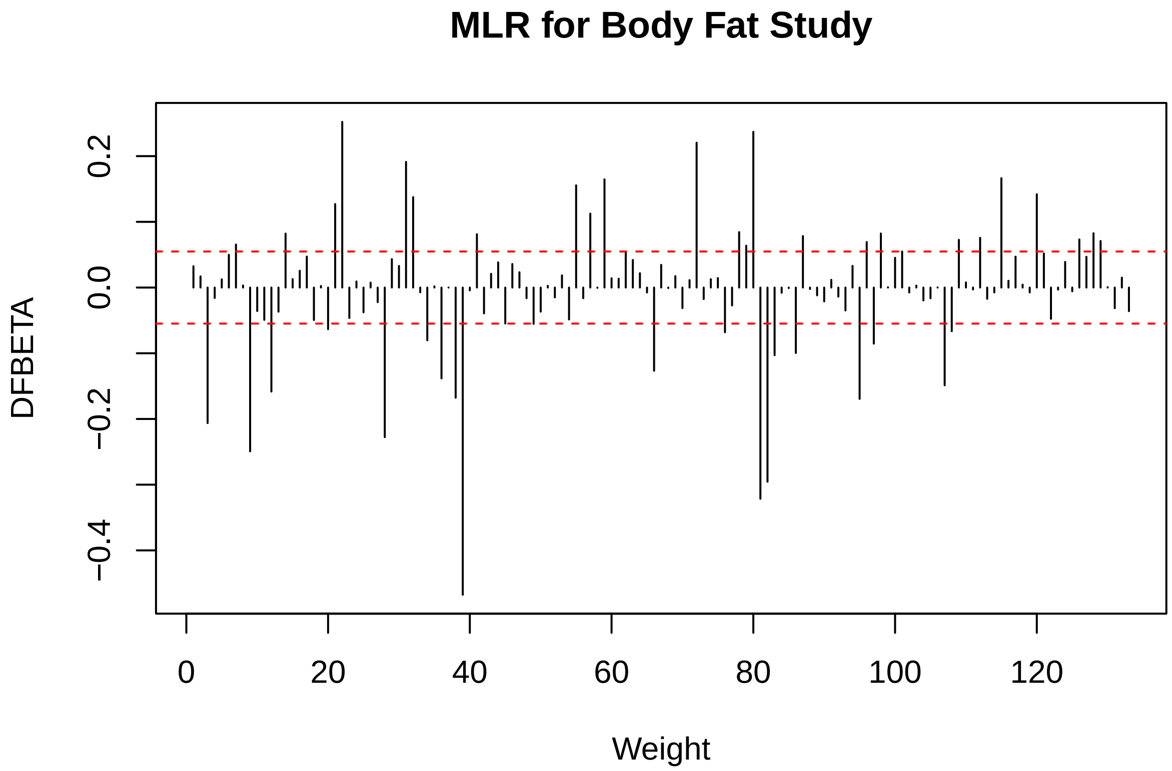
```r
dfb <- dfbetas(best.fat)
# dfb[which(abs(dfb) > 2/sqrt(length(dfb)))]
plot(dfb[,1], type = 'h', main="MLR for Body Fat Study",
ylab="DFBETA", xlab="(Intercept)")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



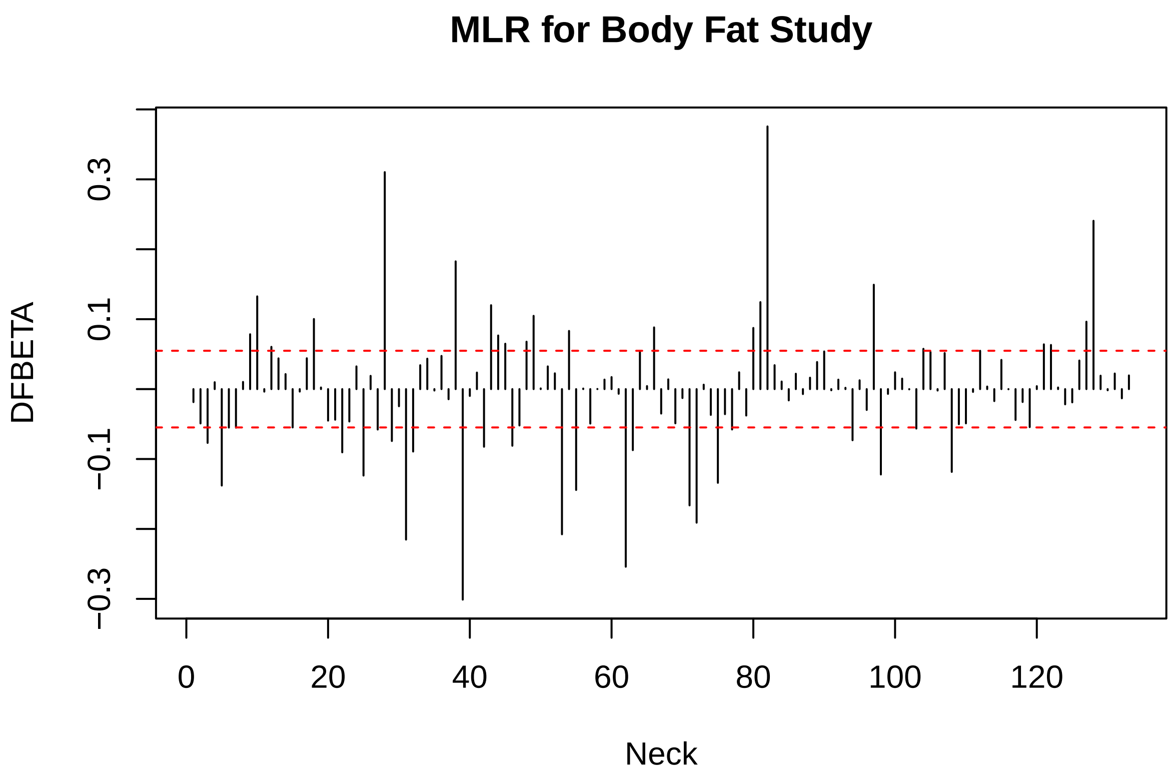
```
# age, weight, neck, abdomen, hip, thigh, ankle, forearm, wrist

# dfb <- dfbetas(best.fat)
# # dfb[which(abs(dfb) > 2/sqrt(length(dfb)))]
# plot(dfb[,1], type = 'h', main="MLR for Body Fat Study",
# ylab="DFBETA", xlab="(Intercept)")
# abline(h=2/sqrt(length(dfb)), col="red", lty=2)
# abline(h=-2/sqrt(length(dfb)), col="red", lty=2)

plot(dfb[,2],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Age")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



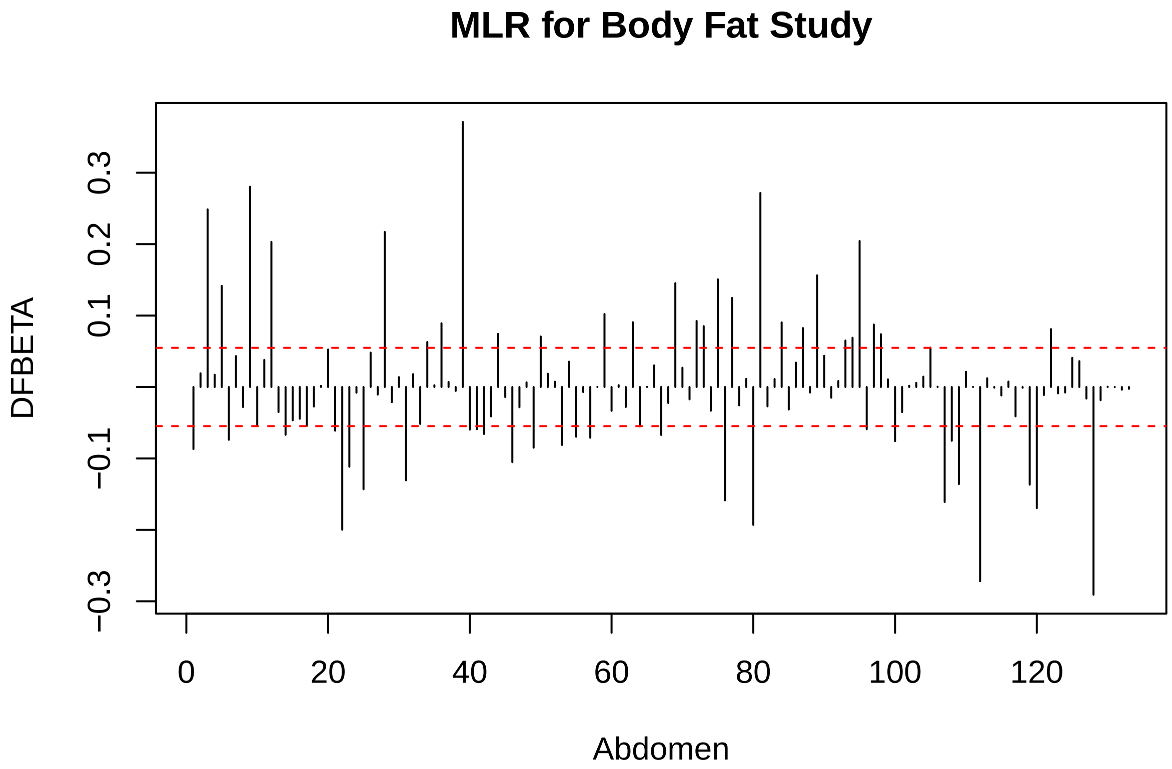
```r
plot(dfb[,3],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Weight")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



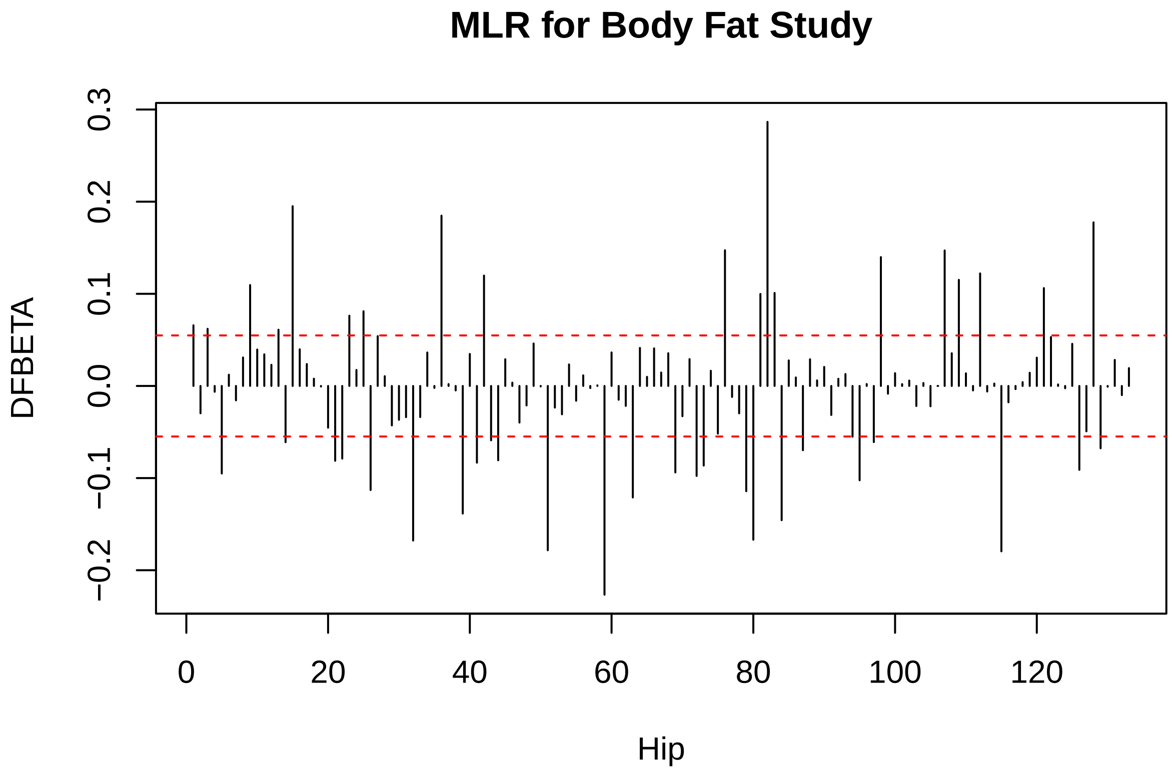
```
plot(dfb[,4],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Neck")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



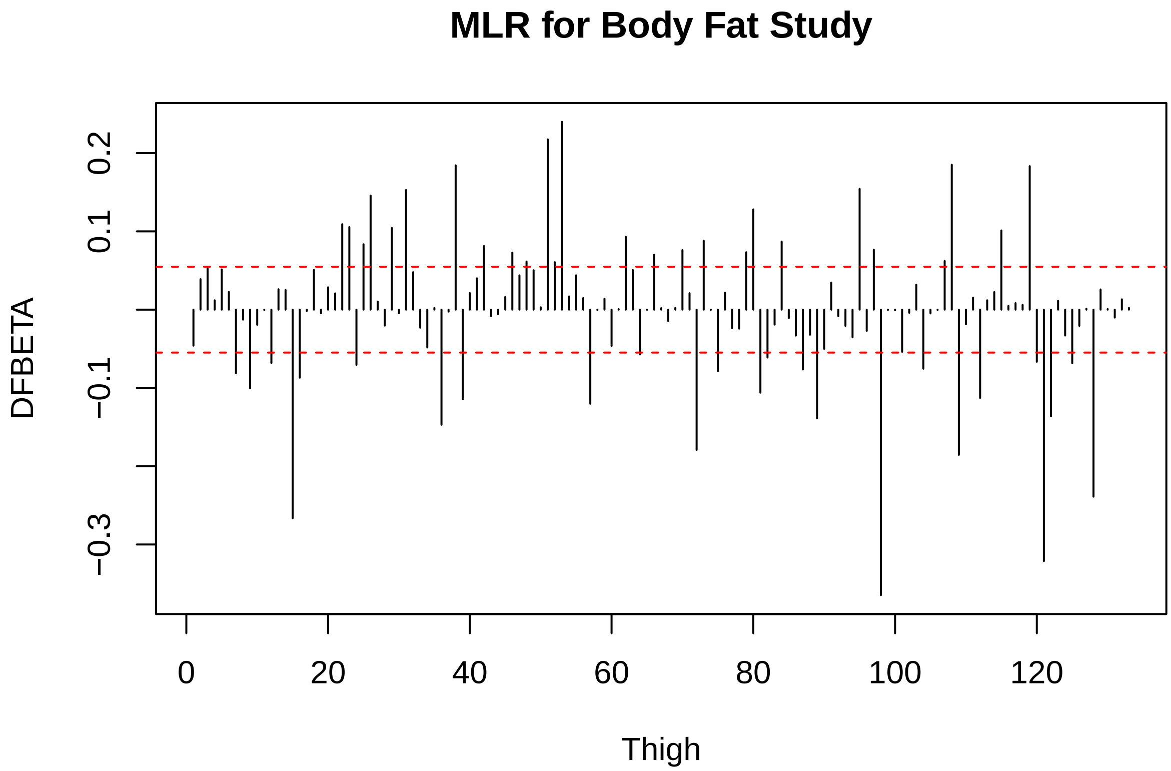
```
plot(dfb[,5],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Abdomen")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



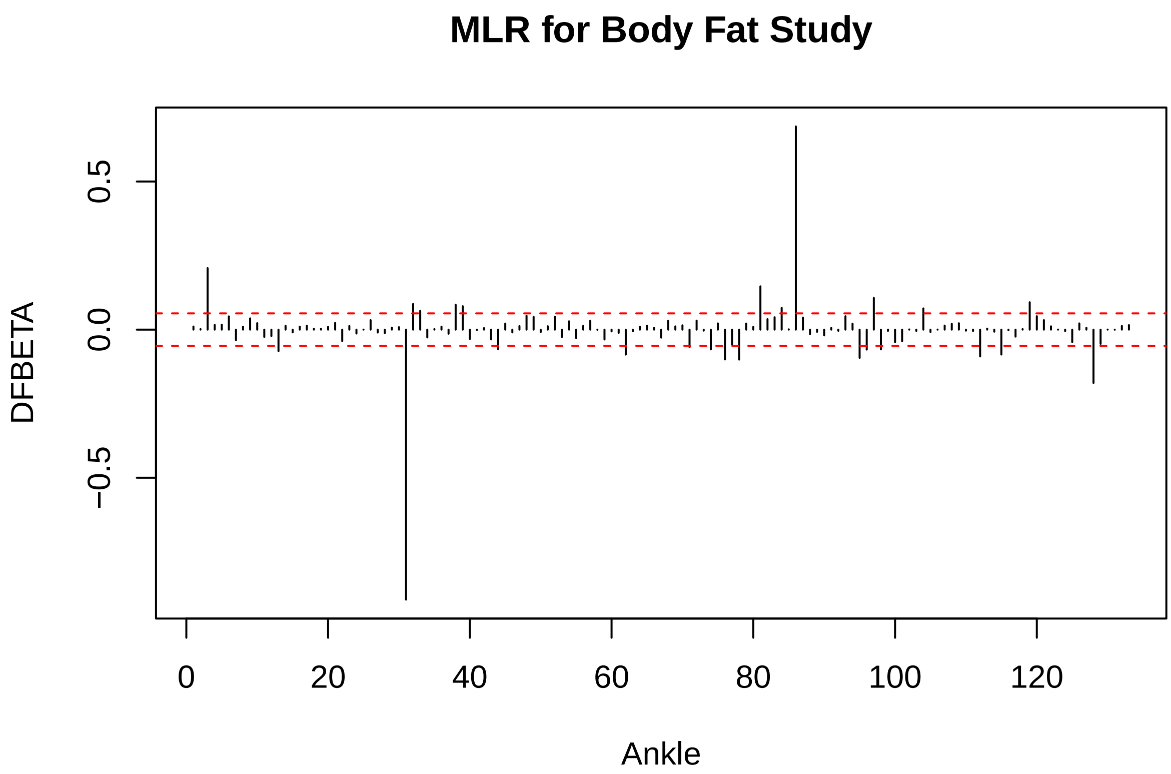
```r
plot(dfb[,6],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Hip")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



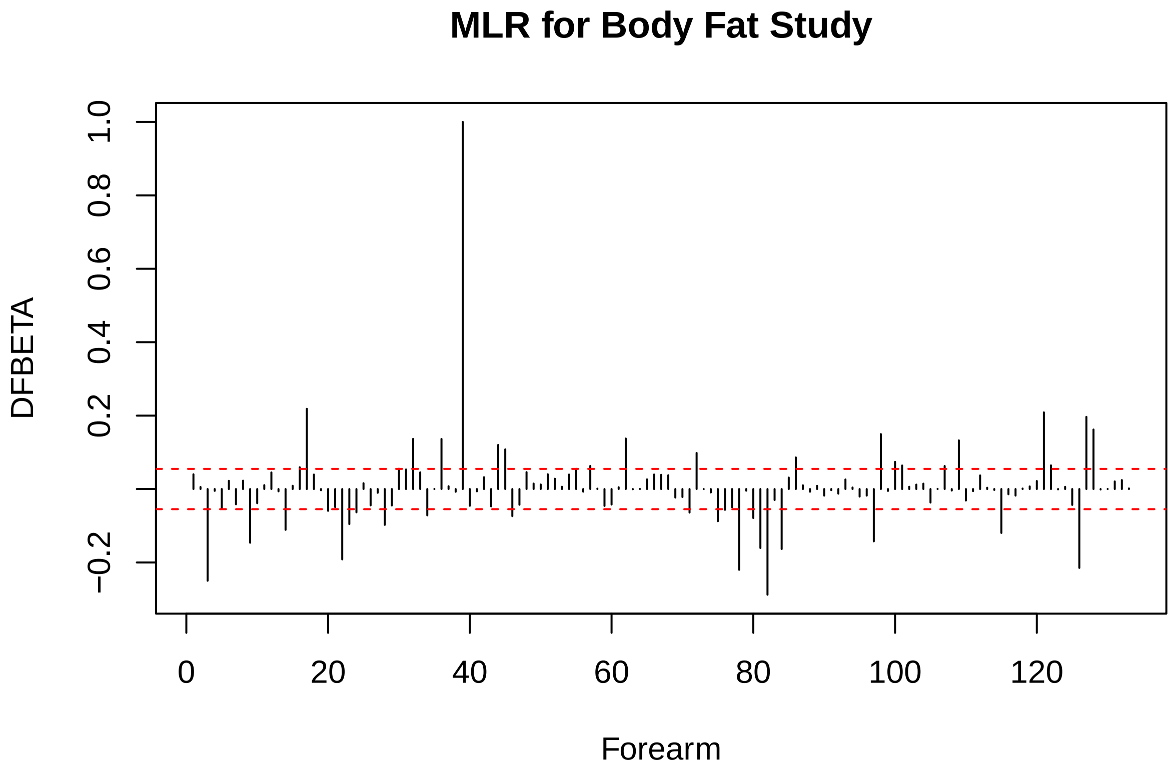
```
plot(dfb[,7],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Thigh")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



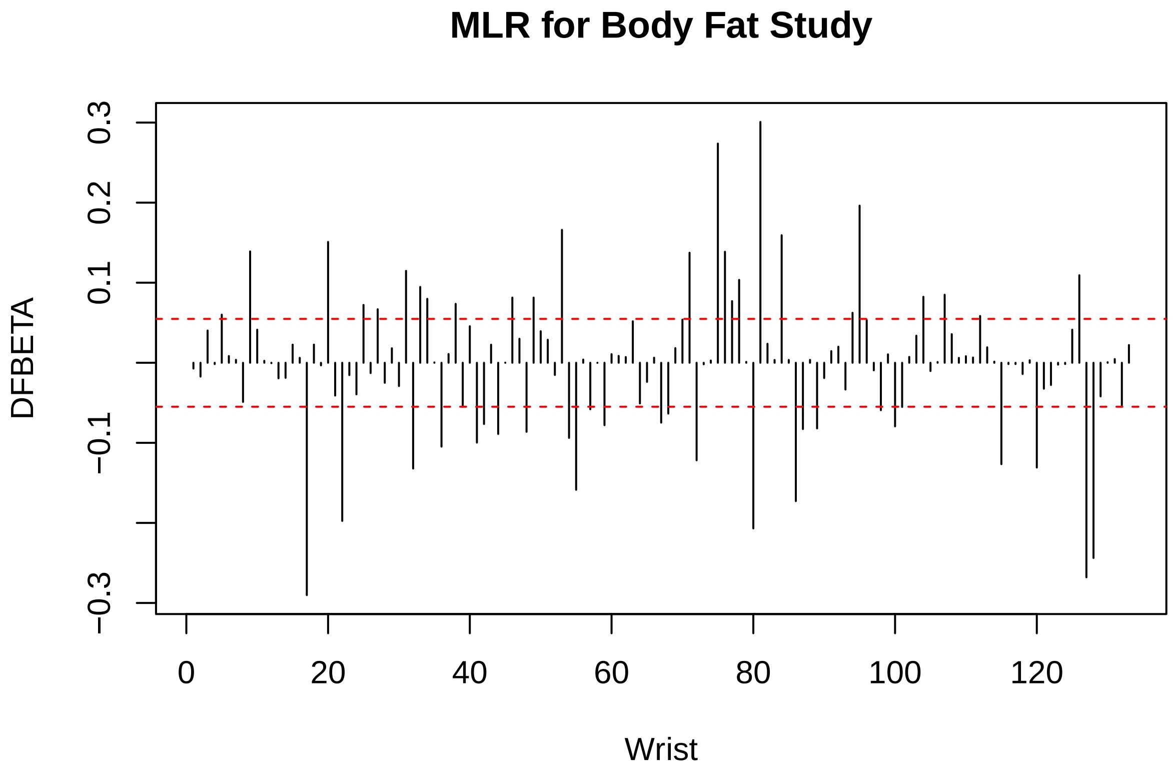
```
plot(dfb[,8],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Ankle")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



```
plot(dfb[,9],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Forearm")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



```
plot(dfb[,10],
     type = 'h',
     main="MLR for Body Fat Study",
     ylab="DFBETA",
     xlab="Wrist")
abline(h=2/sqrt(length(dfb)), col="red", lty=2)
abline(h=-2/sqrt(length(dfb)), col="red", lty=2)
```



There is also the DFBETA method for determining influential points, using the threshold value of $\pm 2/\sqrt{n}$ for each explanatory variable. From (a number of these plots), we also identify potential influential points in the dataset. Overall, we do appear to have a number of outliers, leverage points, and influential points in our data, as shown in a number of plots above. We should consider transforming our data and rerunning a number of our operations to identify a next "best" model and also to diagnose our assumptions for that model.