Department of Statistics

COURSE STATS 330/762

Model Answer to Assignment 3, 2010

1. *Read the data into R. Make a data frame, naming the variables with the names above. Print out the first 10 lines. [5 marks]*

The following code reads in the data, and names the variables. Note that there is no header line in the file.

```
pbf.df =
read.table("C:\\Users\\alee044\\Documents\\Teaching\\330\\assignments\\2010\\Assignme
nt 3\\PBF.txt",header=FALSE)

names(pbf.df) = c("PBF", #Percent body fat using equation, 457/Density - 414.2
"Density", #Density (gm/cm^3)
"Age",     #Age (yrs)
"Weight", # Weight (lbs)
"Height", # Height (inches)
"BMI", # Adiposity index = Weight/Height^2 (kg/m^2)
"Neck", # Neck circumference (cm)
"Chest", # Chest circumference (cm)
"Abdomen", # Abdomen circumference (cm)
"Hip", # Hip circumference (cm)
"Thigh", # Thigh circumference (cm)
"Knee", # Knee circumference (cm)
"Ankle", # Ankle circumference (cm)
"Biceps", # Extended biceps circumference (cm)
"Forearm", # Forearm circumference (cm)
"Wrist") # Wrist circumference (cm) "distal to the styloid processes"
```

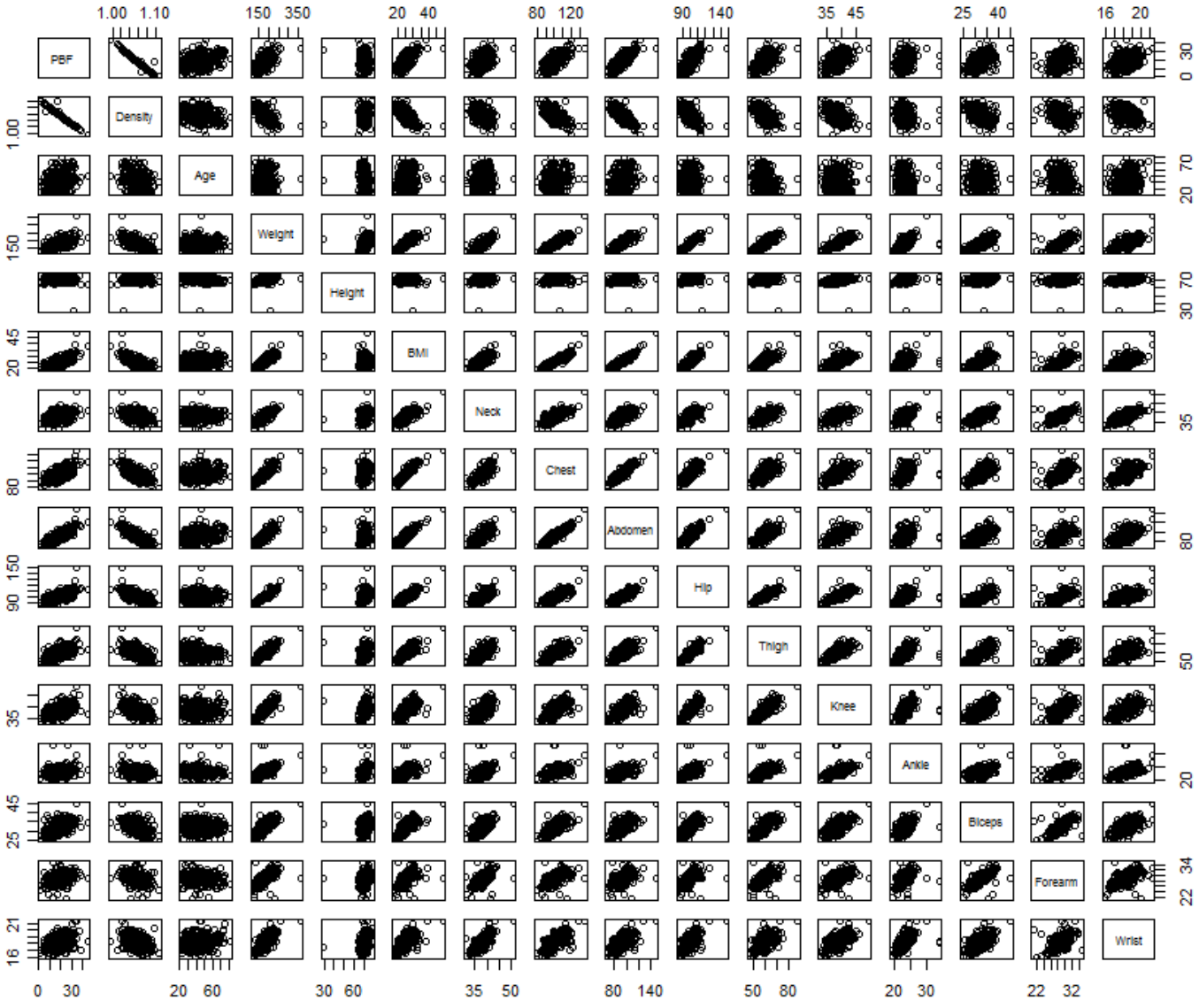
Note that all characters after the # are ignored. The first 10 lines are

```
> pbf.df[1:10,]
    PBF Density Age Weight Height  BMI Neck Chest Abdomen   Hip Thigh Knee Ankle Biceps Forearm Wrist
1  12.6  1.0708  23 154.25  67.75 23.7 36.2  93.1    85.2  94.5  59.0 37.3  21.9   32.0    27.4  17.1
2   6.9  1.0853  22 173.25  72.25 23.4 38.5  93.6    83.0  98.7  58.7 37.3  23.4   30.5    28.9  18.2
3  24.6  1.0414  22 154.00  66.25 24.7 34.0  95.8    87.9  99.2  59.6 38.9  24.0   28.8    25.2  16.6
4  10.9  1.0751  26 184.75  72.25 24.9 37.4 101.8    86.4 101.2  60.1 37.3  22.8   32.4    29.4  18.2
5  27.8  1.0340  24 184.25  71.25 25.6 34.4  97.3   100.0 101.9  63.2 42.2  24.0   32.2    27.7  17.7
6  20.6  1.0502  24 210.25  74.75 26.5 39.0 104.5    94.4 107.8  66.0 42.0  25.6   35.7    30.6  18.8
7  19.0  1.0549  26 181.00  69.75 26.2 36.4 105.1    90.7 100.3  58.4 38.3  22.9   31.9    27.8  17.7
8  12.8  1.0704  25 176.00  72.50 23.6 37.8  99.6    88.5  97.1  60.0 39.4  23.2   30.5    29.0  18.8
9   5.1  1.0900  25 191.00  74.00 24.6 38.1 100.9    82.5  99.9  62.9 38.3  23.8   35.9    31.1  18.2
10 12.0  1.0722  23 198.25  73.50 25.8 42.1  99.6    88.6 104.1  63.1 41.7  25.0   35.6    30.0  19.2
```

2. *I have not changed any values in the original data set, but there are several strange values. Identify these using graphical methods and either correct them or delete the offending observations. (Delete a maximum of 4) . In particular, some of the PBF*

values seem suspect (which ones?) Calculate the volume from the variables Density and Weight. [5 marks]

For a first look, we can calculate a pairs plot:



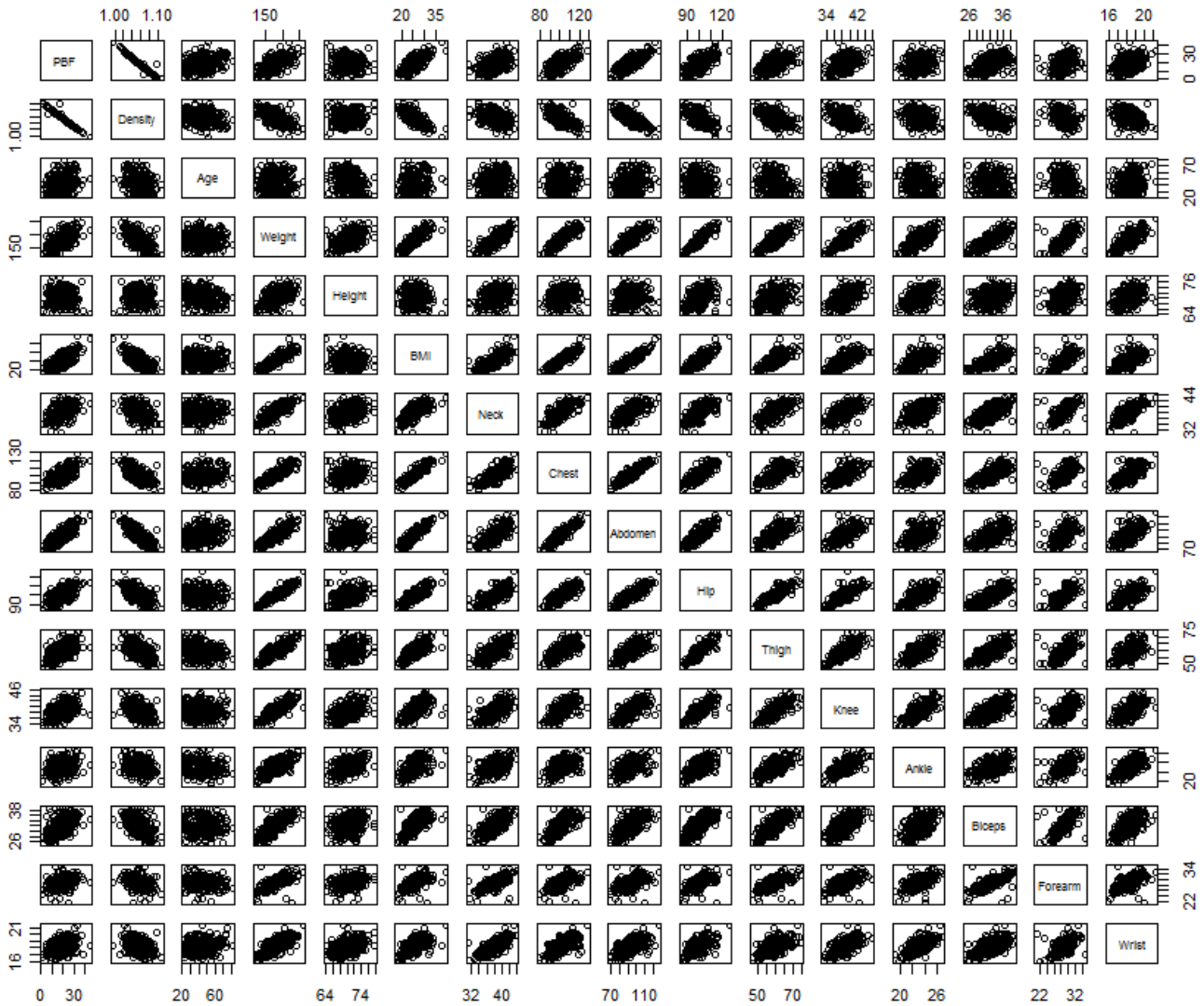
Seems though there are a lot f outliers. To identify them, we can use the function order. This will give the index of the largest and smallest observations for a particular variable:

```
> order(pbf.df$Weight)
  [1] 182  74  45 172 226  50 241  27  29  47 248  49  55 159  53  52 164  23 211 224  75 149 176
 [24] 183 153  72 231  28  76  48  24 217 202  82 177  25 161  67 124 128 151 171  54 191 221   3
 [47]   1 218 220  68  69 239 246 146 197  99  70 154 134  87 235 116  51  26 184 144 210  88  32
 [70] 123 209  30  73 125  79 223 234  77  98  16 195 114  81  46  93 130 106 207 167 137  86 126
 [93] 143 199  71 186 230  33  85 198 135 139 213 110 236 227 185  84 111 200  80 132 156 131 103
[116] 170 215   2 142 102 229 115 233   8  90 174  89 141  91 105 117 160 173 127  78  64 113 201
```

```
[139]  62  21 118  60  92 225 196  13   7  57  31 232 158 112  66  19 204   5  95 163   4 190 122
[162] 136  11 250 179 120  15 138 145  97 240  83 251   9 119 237  36 129  94  63 193 109 214 208
[185]  17 162  38 104 100 133  56  10 101 228 219 245 107 155  22 203 249  58  37 189  59  40 108
[208]  42  14  65 157 121 148 252 147 206  18   6 188  20  44 140 247  12  61 212  43 166  34 165
[231] 216 238 181 150 205  96 242 168 194 175 244 169 222 187 243 180 178 152 192  35  41  39
```

The large weight visible in the pairs plot is observation 39. Similarly the small height is observation 42 (with a height of29,5 inches and a weight of 205 pounds!), the large hip, thigh and knee all 39, the two large ankles 31 and 86.
Deleting these (and the last 2 observations as well) gives a better pairs plot:

```
pairs(pbf.df[-c(31,39,42,86),])
```

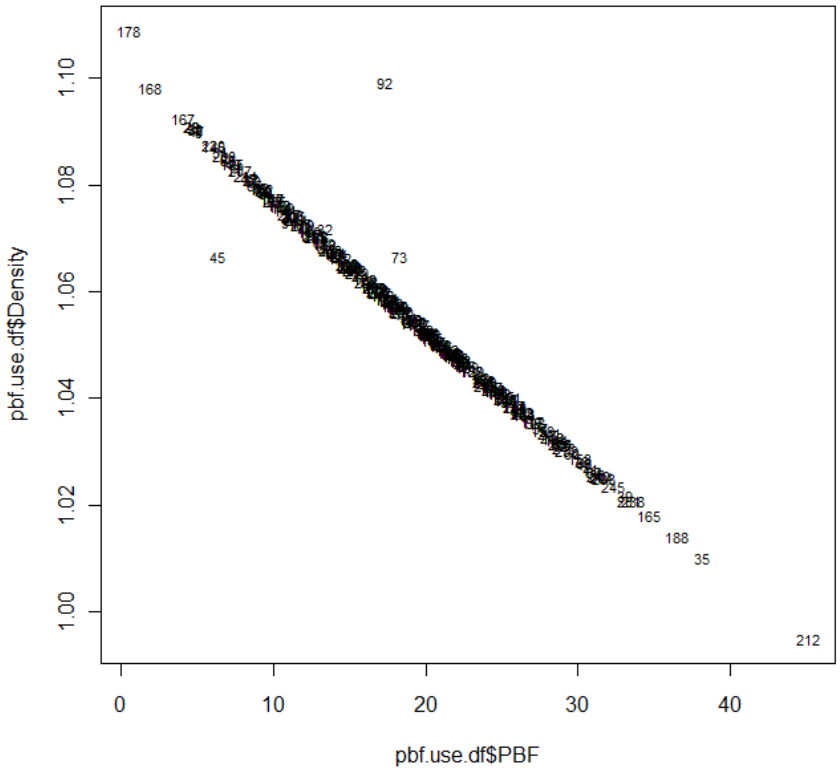


There are some other points liable to have large influence, but we have used up our outlier budget.
Lets work with this reduced data set, plus deleting the last 2 observations:

```
pbf.use.df =  pbf.df[-c(31,39,42,86,251,252),]
```

We will adjust the row labels to be 1 to 248:

To check out the accuracy of the PBF calculation, plot the calculated PBF against the density, using the index number as the plotting symbol:



Seems like 4 points are miscalculated, namely 45, 73, 92, 178.

Since these will not affect rest of the analysis, we won't correct them.

To calculate the volume in litres and add this variable to the data frame, we type

```
Volume = pbf.use.df$Weight*453.59237/ (1000*pbf.use.df$Density)
pbf.use.df = data.frame(pbf.use.df, Volume)
```

3. Develop a model that will predict the volume from the other variables, excluding Density and PBF. You should be able to come up with a model that predicts very well. Points to note: Which variables should be selected? Are transformations indicated? (think Cherry trees). You should potentially consider using all the techniques you have been taught, up to the end of lecture 15. [20 marks]

It seems as though the relationship between Volume and the other variables could well be multiplicative, as it was with the cherry trees. Let's log all the variables, making new, logged variables. This will be necessary for the variable selection methods to work. We will eliminate Density and PBF as they are no longer needed.

Here is a quick way to do this (it works because all the variables are numeric)

```
> log.pbf.df = log(pbf.use.df)[,-c(1,2)]
> newnames = paste("log.", names(pbf.use.df)[-c(1,2)], sep="")
> newnames
[1] "log.Age"     "log.Weight"  "log.Height"  "log.BMI"     "log.Neck"
[6] "log.Chest"   "log.Abdomen" "log.Hip"     "log.Thigh"   "log.Knee"
[11] "log.Ankle"   "log.Biceps"  "log.Forearm" "log.Wrist"   "log.Volume"
> names(log.pbf.df)=newnames
```

There are substantial correlations between the variables, so not all will be needed. To figure out which ones are required, we do some variable selection. First all possible regressions:

```
      rssp sigma2  adjRsq        Cp      AIC      BIC      CV log.Age
1  0.04933  2e-04 0.99244 326.36360 572.3636 579.3743 0.00490       0
2  0.02322  1e-04 0.99643  27.47432 273.4743 283.9903 0.00233       0
3  0.02179  9e-05 0.99663  13.05452 259.0545 273.0759 0.00220       0
4  0.02129  9e-05 0.99670   9.26040 255.2604 272.7871 0.00217       0
5  0.02105  9e-05 0.99672   8.48006 254.4801 275.5120 0.00217       0
6  0.02084  9e-05 0.99674   8.08625 254.0863 278.6236 0.00216       0
7  0.02064  9e-05 0.99676   7.81664 253.8166 281.8593 0.00216       0
8  0.02041  9e-05 0.99678   7.15648 253.1565 284.7045 0.00215       1
9  0.02026  9e-05 0.99679   7.43173 253.4317 288.4850 0.00216       1
10 0.02018  9e-05 0.99679   8.47928 254.4793 293.0379 0.00216       1
11 0.02011  9e-05 0.99679   9.71648 255.7165 297.7805 0.00217       1
12 0.02009  9e-05 0.99678  11.42745 257.4275 302.9968 0.00223       1
13 0.02007  9e-05 0.99677  13.23951 259.2395 308.3141 0.00225       1
14 0.02005  9e-05 0.99676  15.00000 261.0000 313.5800 0.00228       1
   log.Weight log.Height log.BMI log.Neck log.Chest log.Abdomen log.Hip
1           1          0       0        0         0           0       0
2           1          0       0        0         0           1       0
3           1          0       0        0         0           1       0
4           1          1       0        0         0           1       0
5           1          1       0        0         1           1       0
6           1          1       0        1         0           1       0
7           1          1       0        1         1           1       1
8           1          1       0        1         1           1       0
9           1          1       0        1         1           1       1
10          1          1       0        1         1           1       1
11          1          1       0        1         1           1       1
12          1          1       1        1         1           1       1
13          1          1       1        1         1           1       1
14          1          1       1        1         1           1       1
   log.Thigh log.Knee log.Ankle log.Biceps log.Forearm log.Wrist
1          0        0         0          0           0         0
2          0        0         0          0           0         0
3          0        0         0          0           0         1
4          0        0         0          0           0         1
5          0        0         0          0           0         1
6          0        0         0          1           0         1
7          0        0         0          0           0         1
8          0        0         0          1           0         1
9          0        0         0          1           0         1
10         0        0         0          1           1         1
11         1        0         0          1           1         1
12         1        0         0          1           1         1
13         1        1         0          1           1         1
14         1        1         1          1           1         1
```

Either the 4-variable model with log.Weight, log.Height, log.Neck, log.Abdomen, log.Wrist (on the basis of BIC, almost the smallest CV) or the 8-variable model with the above plus log.Age, log,Neck, log.Chest and log.Biceps seems indicated.

Stepwise regression gives

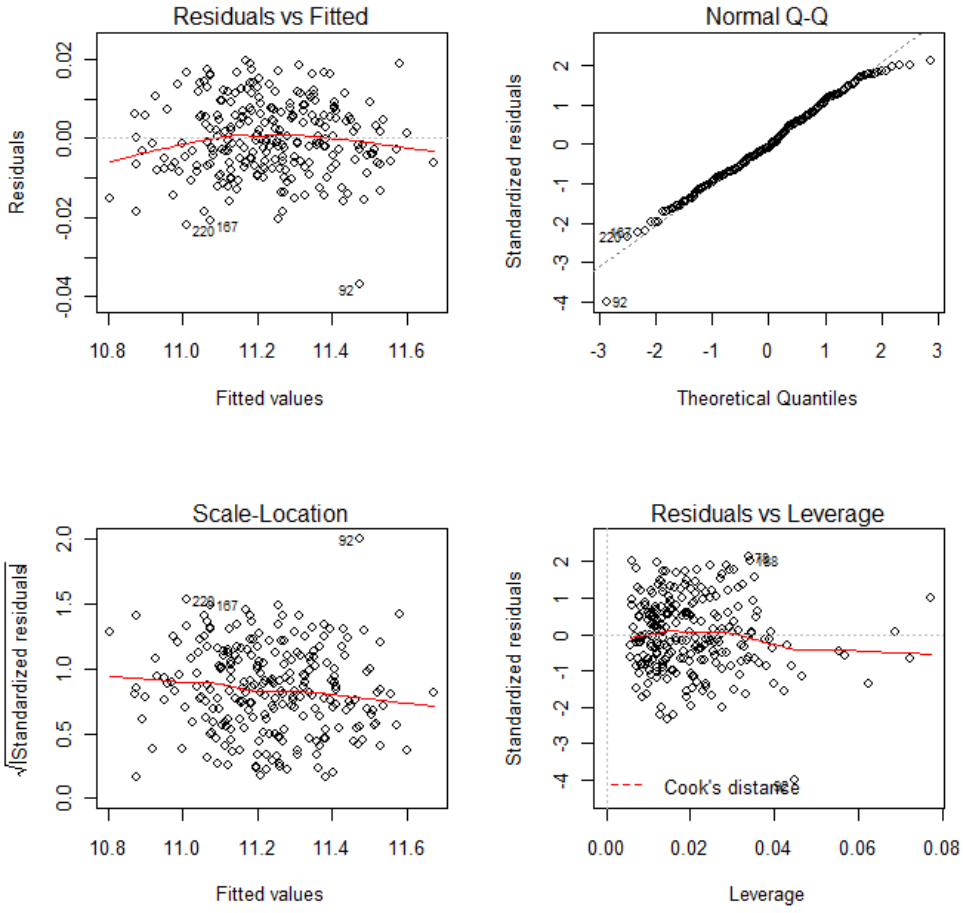
```
model1.lm = lm(log.Volume~., data=log.pbf.df)
null.lm = lm(log.Volume~1, data=log.pbf.df)
step(null.lm, scope=formula(model1.lm), direction="both")

…output not shown
Call:
lm(formula = log.Volume ~ log.Weight + log.Abdomen + log.Wrist +     log.Height + log.Chest +
log.Age + log.Neck + log.Biceps,     data = log.pbf.df)
```

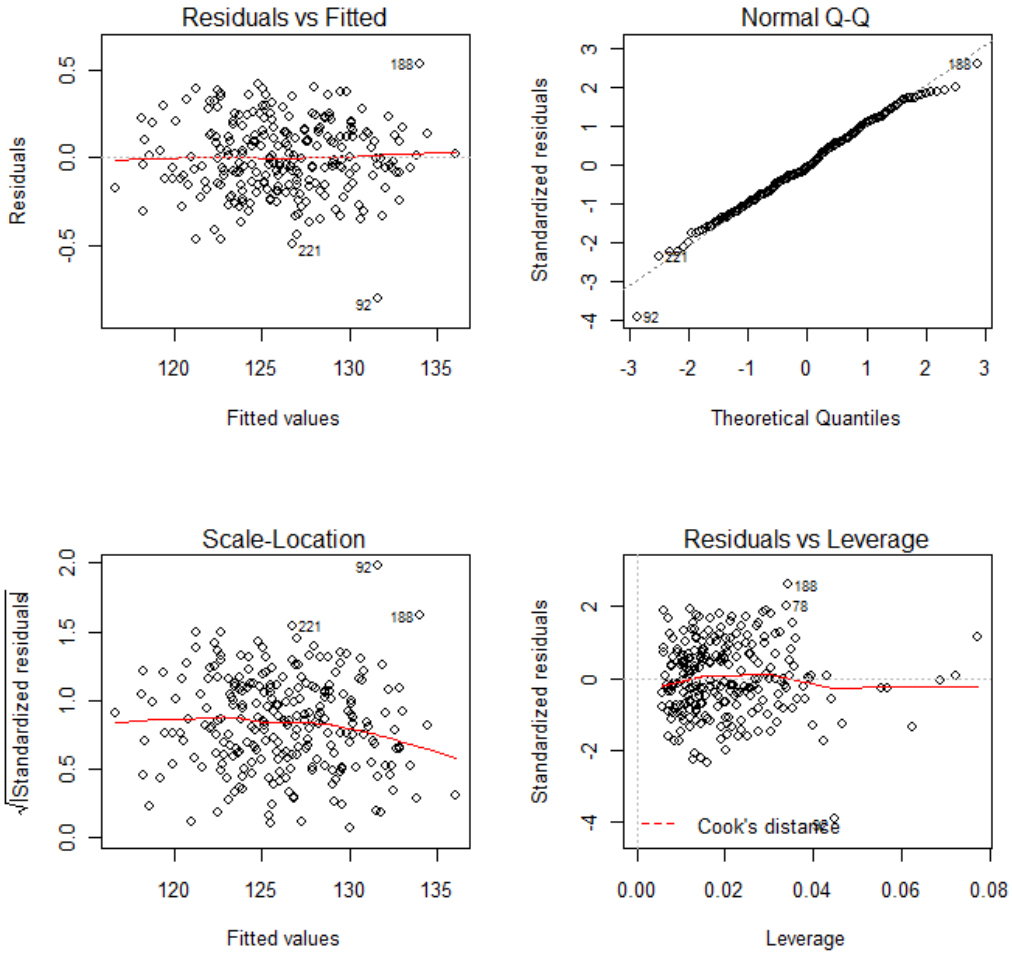
which is the same as the eight-variable model chosen by APR. Both the 4 and 8-variable models should be OK for prediction. In fact the 4 variable model has an R^2 almost as good as the 8-variable model, so we go with the 4-variable model in the interest of simplicity.

Let's subject these models to some checks:

```
> model.6.lm = lm(log.Volume~log.Age + log.Weight + log.Height +
+  log.Neck + log.Abdomen + log.Wrist, data=log.pbf.df)
> par(mfrow=c(2,2))
> plot(model.6.lm)
```



A hint ov curvature is present, and the Boxcox plot (not shown) indicates squaring the response might be a good idea. This leads to the model with plots



which look good apart from the except for an outlier pt 92. This however does not seem to be affecting the coefficients too much, as the influence plot shows. (Cooks distance is OK). Cov ratio indicates point 92 is affecting standard errors. No big outliers. Model seems OK, could use it for prediction. We will explore the effect of point 92 in the predictions.

4. *I have replaced the values of the variables PBF and Density on the last two individuals in the data set with NA's . Using your model, predict the body volume for these two individuals. [10 marks]*

Code for predictions (with and without the high CR point 92:

```
# predictions

# 4.var model, all data

predict.df = log(pbf.df[251:252,-(1:2)])
names(predict.df) = names(log.pbf.df)[-15]
```

```
> exp(sqrt(predict(model.42.lm, predict.df, interval="p")))
         fit      lwr      upr
251 82.87071 81.31958 84.44463
252 90.54701 88.86330 92.25543

# without pt 92
lv2.no92=lv2[-92]
model.42.no92.lm = lm(lv2.no92~ log.Weight + log.Height +
 log.Abdomen + log.Wrist , data=log.pbf.df[-92,])


> exp(sqrt(predict(model.42.no92.lm, predict.df, interval="p")))
                  fit      lwr      upr
251 82.92558 81.41133 84.46152
252 90.62210 88.97804 92.28970
```

The results are very similar. We will go with the last one.

NB: There will be a prize for the best predictions. In the event of a tie, a stochastic mechanism will be used.

Extra question for 762 students

Suppose we logged the volume and the other variables (excluding PBF and Density), and fitted a model to log volume, using the other logged variables. Can you explain why we would not need to include the variable log(BMI) in the model, given the other variables are included?

Since

log(BMI) = log(Weight/Height^2) = log(Weight) – 2log(Height)

there is an exact linear relationship between the logged variables. Thus, if Weight and Height are in the model, adding BMI will not reduce the RSS at all (ie we have perfect collinearity.) If we did try to include it, the software would just ignore it.