Variables lead to Obesity's Model Constructing

2023-05-08

Visualizing the data, Exploratory analysis

```
library(visreg)

Train <- read.table(file='Train.txt', header=TRUE,sep="")

Train[1:10,]
```

```
##    brozek neck chest abdom   hip thigh knee ankle biceps forearm wrist
## 1    12.6 36.2  93.1  85.2  94.5  59.0 37.3  21.9   32.0    27.4  17.1
## 2     6.9 38.5  93.6  83.0  98.7  58.7 37.3  23.4   30.5    28.9  18.2
## 3    10.9 37.4 101.8  86.4 101.2  60.1 37.3  22.8   32.4    29.4  18.2
## 4    27.8 34.4  97.3 100.0 101.9  63.2 42.2  24.0   32.2    27.7  17.7
## 5    19.0 36.4 105.1  90.7 100.3  58.4 38.3  22.9   31.9    27.8  17.7
## 6     5.1 38.1 100.9  82.5  99.9  62.9 38.3  23.8   35.9    31.1  18.2
## 7    12.0 42.1  99.6  88.6 104.1  63.1 41.7  25.0   35.6    30.0  19.2
## 8     7.5 38.5 101.5  83.6  98.2  59.7 39.7  25.2   32.8    29.4  18.5
## 9     8.5 39.4 103.6  90.9 107.7  66.2 39.2  25.9   37.2    30.2  19.0
## 10   20.5 38.4 102.0  91.6 103.9  63.4 38.3  21.5   32.5    28.6  17.7
```

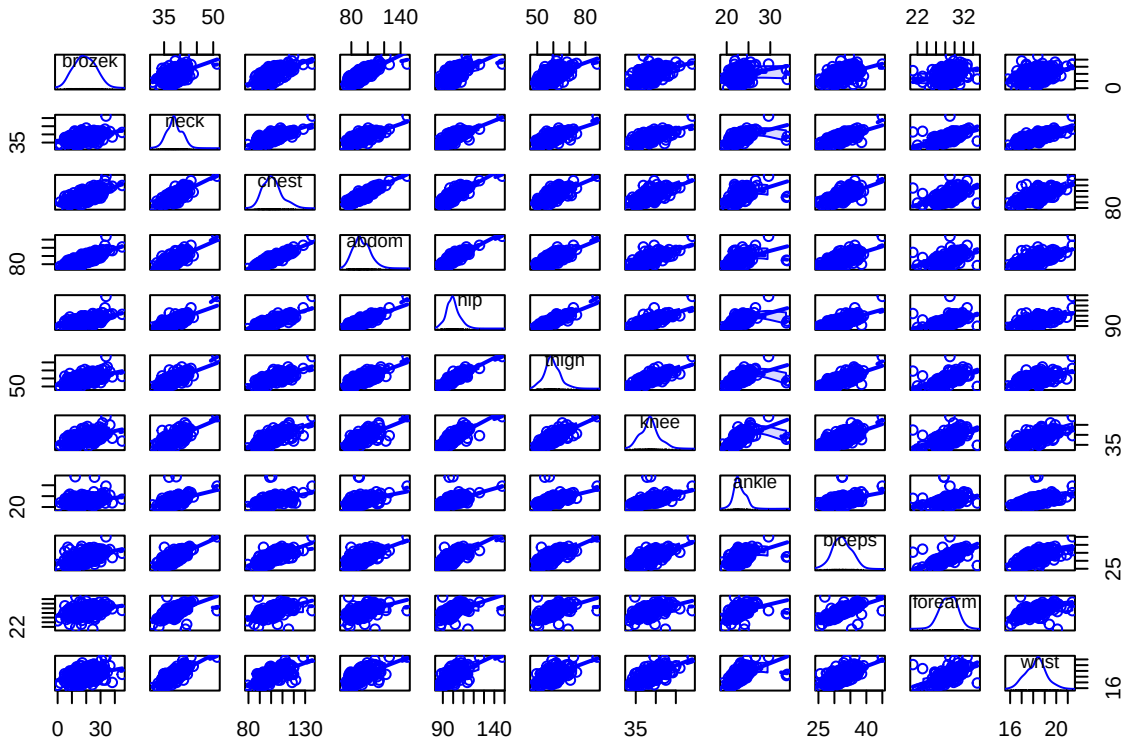
```
str(Train) # Look at the data in details, like numbers of variables and observations
```

```
## 'data.frame':    202 obs. of  11 variables:
##  $ brozek : num  12.6 6.9 10.9 27.8 19 5.1 12 7.5 8.5 20.5 ...
##  $ neck   : num  36.2 38.5 37.4 34.4 36.4 38.1 42.1 38.5 39.4 38.4 ...
##  $ chest  : num  93.1 93.6 101.8 97.3 105.1 ...
##  $ abdom  : num  85.2 83 86.4 100 90.7 82.5 88.6 83.6 90.9 91.6 ...
##  $ hip    : num  94.5 98.7 101.2 101.9 100.3 ...
##  $ thigh  : num  59 58.7 60.1 63.2 58.4 62.9 63.1 59.7 66.2 63.4 ...
##  $ knee   : num  37.3 37.3 37.3 42.2 38.3 38.3 41.7 39.7 39.2 38.3 ...
##  $ ankle  : num  21.9 23.4 22.8 24 22.9 23.8 25 25.2 25.9 21.5 ...
##  $ biceps : num  32 30.5 32.4 32.2 31.9 35.9 35.6 32.8 37.2 32.5 ...
##  $ forearm: num  27.4 28.9 29.4 27.7 27.8 31.1 30 29.4 30.2 28.6 ...
##  $ wrist  : num  17.1 18.2 18.2 17.7 17.7 18.2 19.2 18.5 19 17.7 ...
```

```
library(car)
```

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

```
scatterplotMatrix(Train) ## gives slightly more info
```



The data has 11 variables with 202 observations each. The body fat measurement is the variable brozek (which refers to Brozeks equation for body fat content). The remaining 10 variables are examing the circumferences, in centimeters, of neck, chest, abdom, hip, thigh, knee, ankle, biceps, forearm and wrist. Besides, they are all continuous variables.

The matrix plot shows an overall trends of the relationships between two different variables, it's quite evidental to jugde that all the 10 variables are positively correlated with brozek if ignore the other factors.

Model selection

We want to derive certain models such that through which the value of body fat content (brozek) can be predicted using the other 10 measurements(neck, chest, abadom, hip, thigh, knee, ankle, biceps, forearm, wrist). For the following parts, we assume some models and present them in a nice way. Based on multivariate informations, we would select the best ones among them.

Let's start by fitting a linear model first.

```
fit <- lm(brozek~., data=Train)
coef(fit)
```

```
## (Intercept)        neck       chest       abdom         hip       thigh
##  7.52875822 -0.70585562 -0.09336572  0.98088194 -0.35204966  0.01854230
##        knee       ankle      biceps     forearm       wrist
## -0.10552572 -0.09492463  0.15117121  0.37586590 -1.01543063
```

This is so called "full model" because it contains all the variables with respect to the targeted one(brozek), and we want to know more details about this model.

```
summary(fit)
```

```
##
## Call:
## lm(formula = brozek ~ ., data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2050 -2.6985 -0.2738  2.8135 10.3088
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.52876    6.82825   1.103  0.27159
## neck        -0.70586    0.24230  -2.913  0.00400 **
## chest       -0.09337    0.09177  -1.017  0.31028
## abdom        0.98088    0.07750  12.656  < 2e-16 ***
## hip         -0.35205    0.12741  -2.763  0.00629 **
## thigh        0.01854    0.14111   0.131  0.89560
## knee        -0.10553    0.22432  -0.470  0.63858
## ankle       -0.09492    0.21145  -0.449  0.65400
## biceps       0.15117    0.17110   0.884  0.37805
## forearm      0.37587    0.19598   1.918  0.05662 .
## wrist       -1.01543    0.51112  -1.987  0.04839 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.042 on 191 degrees of freedom
## Multiple R-squared:  0.7451, Adjusted R-squared:  0.7317
## F-statistic: 55.83 on 10 and 191 DF,  p-value: < 2.2e-16
```

Afterwards, we will apply the two automatic variable selection procedures(best subsets regression and stepwise regression) in R to wisely select the models.

Stepwise Regression

```
fit_step1 <- step(fit)
```

```
## Start:  AIC=574.98
## brozek ~ neck + chest + abdom + hip + thigh + knee + ankle +
##     biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - thigh    1      0.28 3120.9 573.00
## - ankle    1      3.29 3123.9 573.19
## - knee     1      3.62 3124.2 573.21
## - biceps   1     12.75 3133.4 573.80
## - chest    1     16.91 3137.5 574.07
## <none>                 3120.6 574.98
```

```
## - forearm  1     60.10 3180.7 576.83
## - wrist    1     64.48 3185.1 577.11
## - hip      1    124.74 3245.3 580.90
## - neck     1    138.66 3259.3 581.76
## - abdom    1   2616.91 5737.5 696.00
##
## Step:  AIC=573
## brozek ~ neck + chest + abdom + hip + knee + ankle + biceps +
##     forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - ankle    1      3.23 3124.1 571.21
## - knee     1      3.36 3124.3 571.21
## - biceps   1     15.26 3136.2 571.98
## - chest    1     18.23 3139.1 572.17
## <none>                 3120.9 573.00
## - forearm  1     61.44 3182.3 574.93
## - wrist    1     68.06 3188.9 575.35
## - neck     1    138.38 3259.3 579.76
## - hip      1    191.79 3312.7 583.04
## - abdom    1   2620.99 5741.9 694.15
##
## Step:  AIC=571.21
## brozek ~ neck + chest + abdom + hip + knee + biceps + forearm +
##     wrist
##
##           Df Sum of Sq    RSS    AIC
## - knee     1      5.34 3129.5 569.55
## - biceps   1     15.12 3139.2 570.18
## - chest    1     19.10 3143.2 570.44
## <none>                 3124.1 571.21
## - forearm  1     61.47 3185.6 573.14
## - wrist    1     77.16 3201.3 574.13
## - neck     1    138.91 3263.0 577.99
## - hip      1    205.37 3329.5 582.07
## - abdom    1   2716.36 5840.5 695.59
##
## Step:  AIC=569.55
## brozek ~ neck + chest + abdom + hip + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - biceps   1     13.88 3143.3 568.44
## - chest    1     18.26 3147.7 568.73
## <none>                 3129.5 569.55
## - forearm  1     59.27 3188.7 571.34
## - wrist    1     94.06 3223.5 573.53
## - neck     1    136.76 3266.2 576.19
## - hip      1    286.25 3415.7 585.23
## - abdom    1   2713.09 5842.5 693.66
##
## Step:  AIC=568.44
## brozek ~ neck + chest + abdom + hip + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
```

```
## - chest    1     15.24 3158.6 567.42
## <none>                 3143.3 568.44
## - wrist    1     87.88 3231.2 572.01
## - forearm  1     90.23 3233.6 572.16
## - neck     1    125.63 3269.0 574.36
## - hip      1    273.78 3417.1 583.31
## - abdom    1   2700.46 5843.8 691.70
##
## Step:  AIC=567.42
## brozek ~ neck + abdom + hip + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## <none>                 3158.6 567.42
## - forearm  1      78.9 3237.5 570.41
## - wrist    1      96.9 3255.5 571.52
## - neck     1     139.3 3297.9 574.14
## - hip      1     285.9 3444.5 582.93
## - abdom    1    4233.4 7392.0 737.18
```

In this process, we have 6 models in total. From the definition of stepwise regression and 10 variables need to be examined, we know that there would be 11 models in total in fact; thus, we need to apply another step analysis below. Anyway, from the above models we derived, we judge them by a criteria named "Akaike information criterion"(AIC), it's defined to be AIC = -2log(L) + 2p where log(L) is the log-likelihood calculated at the maximized likelihood estimate of beta and sigma. Besides, the method we apply here is to judge the value of AIC between two different models and the model with less AIC value is preferred. To this end, the model(brozek ~ neck + abdom + hip + forearm + wrist) is preferred since it obtains the lowest AIC value(567.42).

```
fit0 <- lm(brozek~1, data=Train)
fit_step2 <- step(fit0, scope = brozek~neck + chest + abdom + hip + thigh + knee + ankle +  biceps + fo
```

```
## Start:  AIC=831.07
## brozek ~ 1
##
##           Df Sum of Sq     RSS    AIC
## + abdom    1    8141.3  4100.3 612.13
## + chest    1    6055.6  6186.0 695.20
## + hip      1    4686.8  7554.8 735.58
## + thigh    1    3675.0  8566.5 760.97
## + neck     1    3187.2  9054.4 772.15
## + knee     1    3157.4  9084.2 772.82
## + biceps   1    2778.7  9462.9 781.07
## + wrist    1    1637.3 10604.3 804.07
## + forearm  1    1604.9 10636.7 804.69
## + ankle    1     758.3 11483.3 820.16
## <none>                 12241.6 831.07
##
## Step:  AIC=612.13
## brozek ~ abdom
##
##           Df Sum of Sq     RSS    AIC
## + neck     1     498.4  3602.0 587.95
## + hip      1     496.7  3603.6 588.05
```

```
## + wrist    1     450.4  3649.9 590.63
## + knee     1     274.0  3826.3 600.16
## + ankle    1     231.9  3868.4 602.37
## + thigh    1     226.1  3874.2 602.68
## + chest    1     146.5  3953.8 606.78
## + biceps   1     123.9  3976.4 607.93
## <none>                  4100.3 612.13
## + forearm  1      21.5  4078.8 613.07
## - abdom    1    8141.3 12241.6 831.07
##
## Step:  AIC=587.95
## brozek ~ abdom + neck
##
##           Df Sum of Sq    RSS    AIC
## + hip      1     292.2 3309.7 572.86
## + knee     1     122.0 3479.9 582.99
## + wrist    1     108.6 3493.3 583.77
## + thigh    1      93.9 3508.0 584.62
## + ankle    1      90.9 3511.0 584.79
## <none>                 3602.0 587.95
## + chest    1      26.2 3575.7 588.48
## + forearm  1      25.3 3576.6 588.53
## + biceps   1       1.2 3600.8 589.89
## - neck     1     498.4 4100.3 612.13
## - abdom    1    5452.5 9054.4 772.15
##
## Step:  AIC=572.86
## brozek ~ abdom + neck + hip
##
##           Df Sum of Sq    RSS    AIC
## + wrist    1      72.2 3237.5 570.41
## + forearm  1      54.3 3255.5 571.52
## <none>                 3309.7 572.86
## + biceps   1      23.0 3286.7 573.45
## + ankle    1      18.4 3291.4 573.74
## + thigh    1      15.0 3294.7 573.94
## + knee     1      11.3 3298.5 574.17
## + chest    1       9.8 3299.9 574.26
## - hip      1     292.2 3602.0 587.95
## - neck     1     293.9 3603.6 588.05
## - abdom    1    4201.5 7511.2 736.41
##
## Step:  AIC=570.41
## brozek ~ abdom + neck + hip + wrist
##
##           Df Sum of Sq    RSS    AIC
## + forearm  1      78.9 3158.6 567.42
## + biceps   1      39.2 3198.3 569.94
## <none>                 3237.5 570.41
## + thigh    1      10.9 3226.6 571.72
## + ankle    1       4.1 3233.4 572.15
## + chest    1       3.9 3233.6 572.16
## + knee     1       1.1 3236.4 572.34
## - wrist    1      72.2 3309.7 572.86
```

```
## - neck     1      97.9 3335.4 574.42
## - hip      1     255.8 3493.3 583.77
## - abdom    1    4174.3 7411.8 735.72
##
## Step:  AIC=567.42
## brozek ~ abdom + neck + hip + wrist + forearm
##
##           Df Sum of Sq    RSS    AIC
## <none>                 3158.6 567.42
## + chest    1      15.2 3143.3 568.44
## + biceps   1      10.9 3147.7 568.73
## + ankle    1       5.5 3153.1 569.07
## + knee     1       3.6 3155.0 569.19
## + thigh    1       2.8 3155.8 569.24
## - forearm  1      78.9 3237.5 570.41
## - wrist    1      96.9 3255.5 571.52
## - neck     1     139.3 3297.9 574.14
## - hip      1     285.9 3444.5 582.93
## - abdom    1    4233.4 7392.0 737.18
```

After two step regression simulation(the first one is called backward regression and the second one is forward regression in fact), similarlly, we compare the AIC values of these models. Suprisingly, we notice that the model (brozek ~ abdom + neck + hip + wrist + forearm) would be the optimal choice since it obsess the lowest value of AIC of the 11 models.

Best Subsets Regression

By definition, there are 2^10(=1024) models to test.

```
library(leaps)

a <- regsubsets(brozek~., data=Train)
summary.out <- summary(a)
summary.out
```

```
## Subset selection object
## Call: regsubsets.formula(brozek ~ ., data = Train)
## 10 Variables  (and intercept)
##         Forced in Forced out
## neck        FALSE      FALSE
## chest       FALSE      FALSE
## abdom       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
##          neck chest abdom hip thigh knee ankle biceps forearm wrist
## 1  ( 1 ) " "  " "   "*"   " " " " " "   " "   " "    " "     " "
```

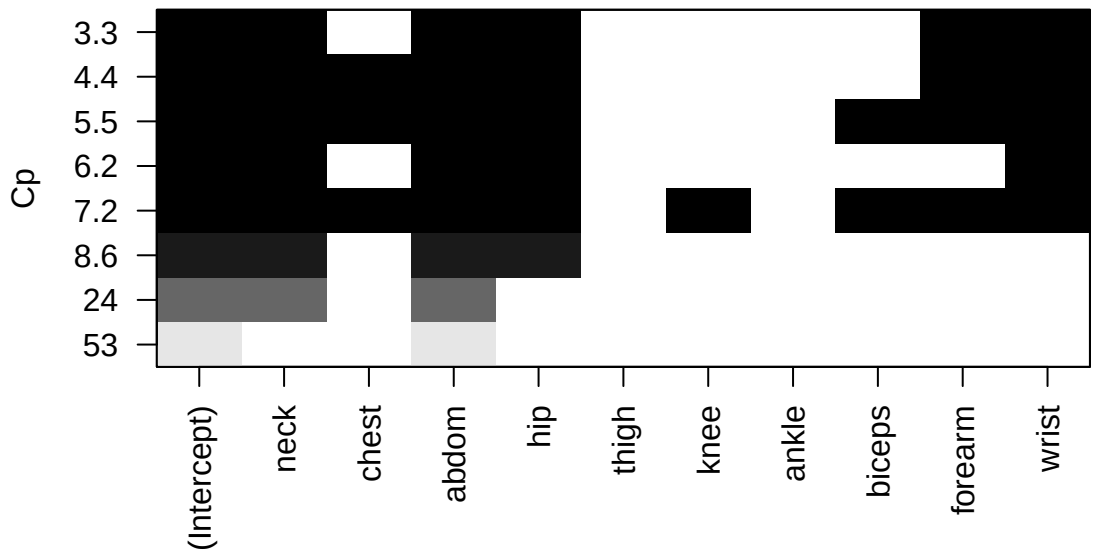
```
## 2  ( 1 ) "*"  " "  "*"  " " " "  " "  " "  " "  " "  " "
## 3  ( 1 ) "*"  " "  "*"  "*" " "  " "  " "  " "  " "  " "
## 4  ( 1 ) "*"  " "  "*"  "*" " "  " "  " "  " "  " "  "*"
## 5  ( 1 ) "*"  " "  "*"  "*" " "  " "  " "  " "  "*"  "*"
## 6  ( 1 ) "*"  "*"  "*"  "*" " "  " "  " "  " "  "*"  "*"
## 7  ( 1 ) "*"  "*"  "*"  "*" " "  " "  " "  "*"  "*"  "*"
## 8  ( 1 ) "*"  "*"  "*"  "*" " "  "*"  " "  "*"  "*"  "*"
```

```
summary.out$cp
```

```
## [1] 52.964732 24.461290  8.574988  6.153231  3.324409  4.391811  5.542002
## [8]  7.215071
```

Based on the value of Cp(Mallow's Cp) which can be automaticly derived by R, and the judging criteria is that the less is preferred.

```
plot(a, scale="Cp")
```



From the plot, we can see that the best model is (brozek ~ abdom + neck + hip + wrist + forearm) since it has the lowest Cp.

Based on the analysis of two different method, it reaches a consensus that the model(brozek ~ abdom + neck + hip + wrist + forearm) is the best among all.

Consequently, we fit the desired model into fit.desired. brozek = intercept + beta1 * abdom + beta2 * neck + beta3 * hip + beta4*wrist* + *beta5*forearm + epislon

```
fit.desired <- lm(brozek ~ abdom + neck + hip + wrist + forearm, data=Train)
summary(fit.desired)
```

```
##
## Call:
## lm(formula = brozek ~ abdom + neck + hip + wrist + forearm, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1901 -2.7695 -0.2035  2.9381 10.1603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.26055    6.26699   0.680  0.49741
## abdom        0.92918    0.05733  16.208  < 2e-16 ***
## neck        -0.68909    0.23436  -2.940  0.00367 **
## hip         -0.35567    0.08444  -4.212 3.85e-05 ***
## wrist       -1.14862    0.46850  -2.452  0.01509 *
## forearm      0.39413    0.17813   2.213  0.02808 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.014 on 196 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.7354
## F-statistic: 112.7 on 5 and 196 DF,  p-value: < 2.2e-16
```

Model Checking and Validation

First of all, from the summary of the fit.desired which is the desired matrix we selected, it returns R-squared with value 0.742, and adjusted R-squared 0.7354, which validates a strong relationship between the variables we have in the model with the targeted value, brozek. That is to say, our model is a nice configuration.

Moreover, we analysis in the way of hypothesis test: H0: beta = 0 where beta = (beta1, beta2, beta3, beta4, beta5) H1: beta != 0 and we take level of confidence to be 95%.

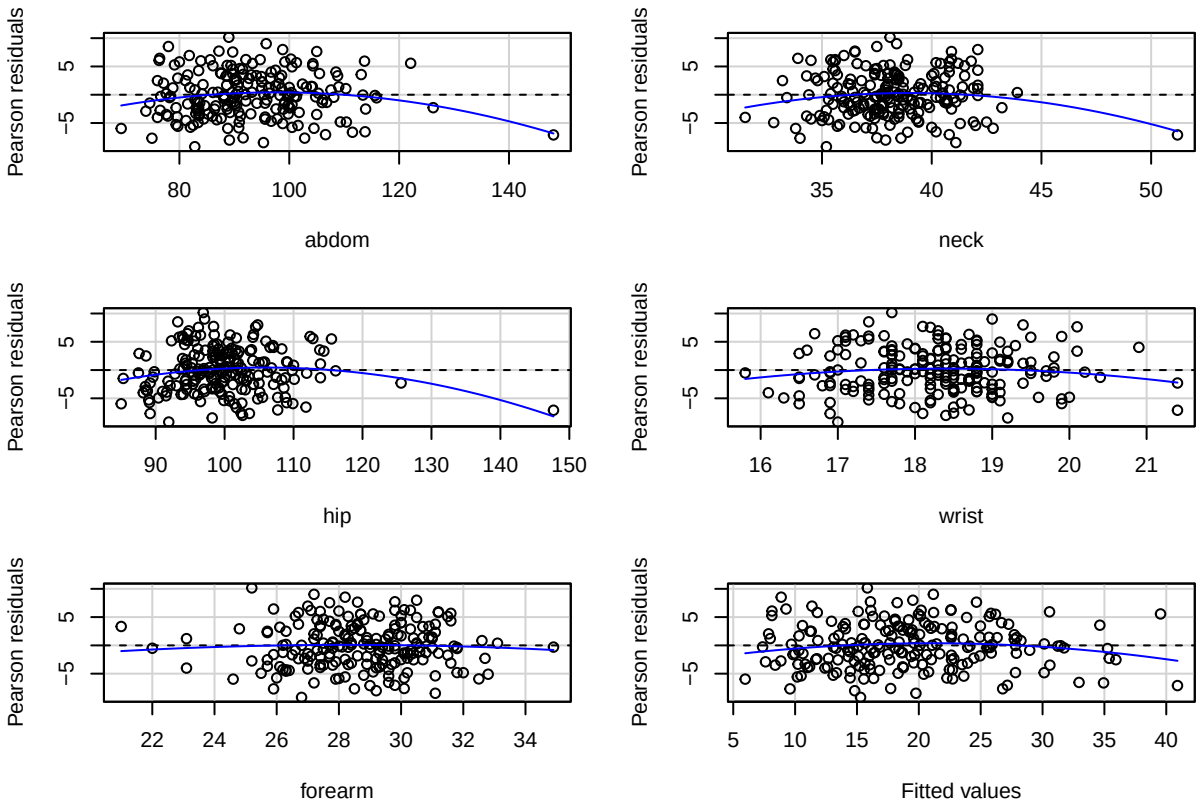
From the p-value of the t-distribution test of those variables (2e-16, 0.00367, 3.85e-05, ect.), they are all relatively small when we set the level of confidence to be 95%. In this way, all 5 variables all at least not weakly correlated to brozek though t-test loses some generality like cannot judge the effect of one variable if it is embedded in a combination with another powerful variable.

From the perspective of F-statistics, the model returns a p-value of 2.2e-16 which is way less than even 1%, that is to say, we have strong evidence against H0.

Residual Plot

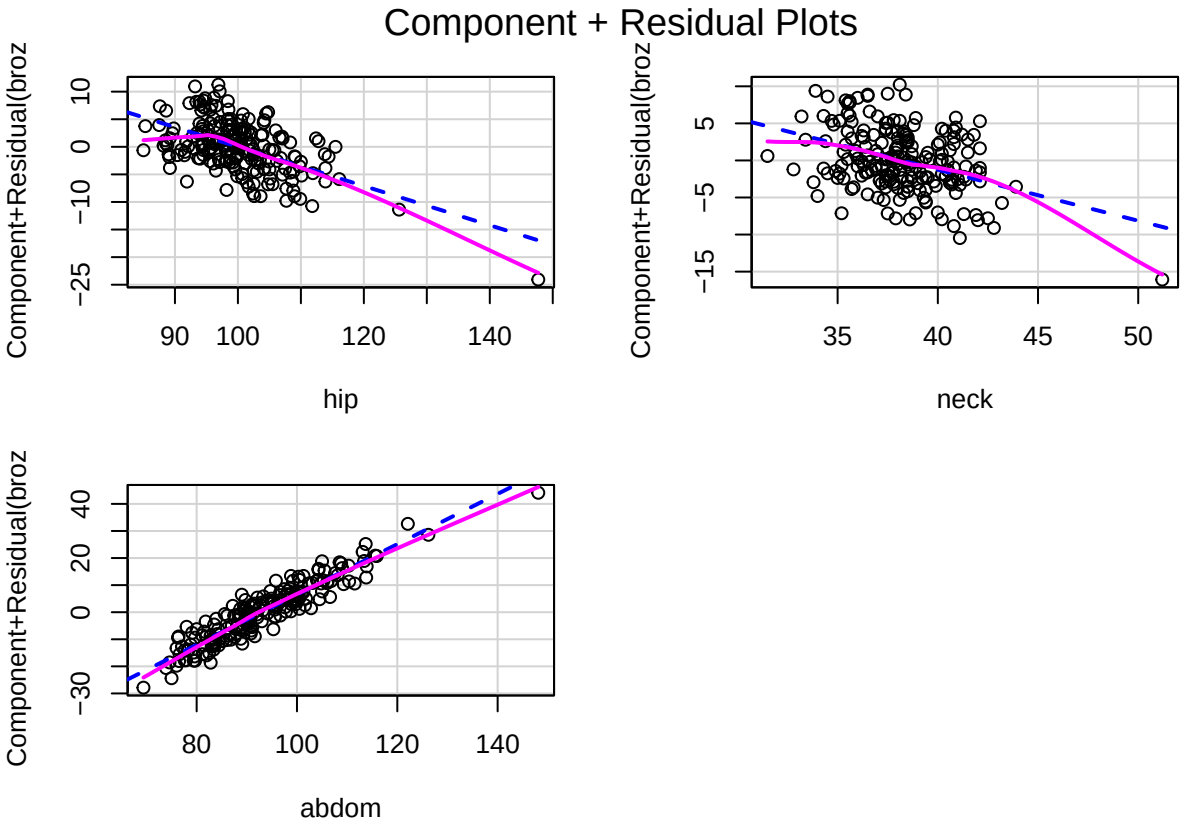
We want to examine our model by examing whether the assumption that the random errors are satisfied: 1. uncorrelated 2. have equal variance 3. have zero mean 4. The errors are normally distributed

```
residualPlots(fit.desired, tests=FALSE)
```

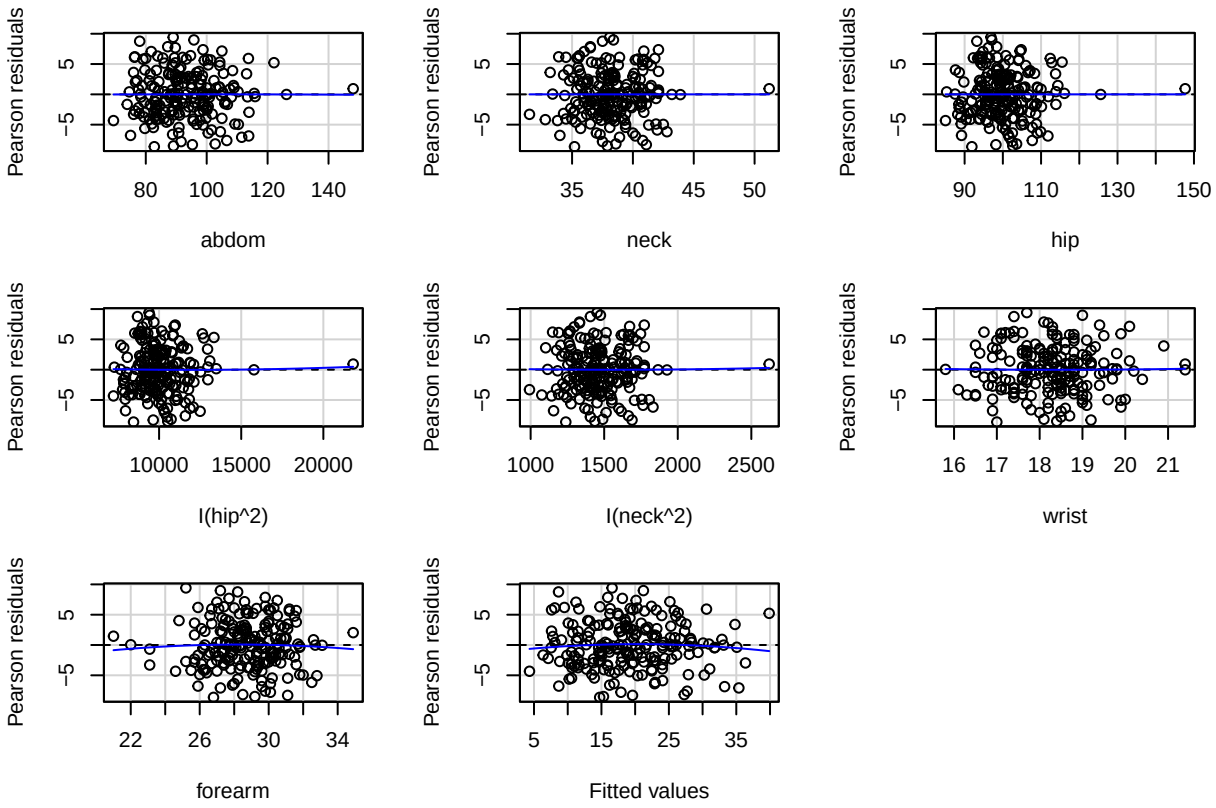


By observing the Fitted Values vs. Pearson residual, we can spot that the distribution is muck alike a null plot (a band of points with no discernible trend between the residual and the fitted value), which is highly acceptable meaning that our regression model is correct. But there are also some variations for certain variables (hip, neck, abdom). Thus, we may derive a better model from this one, e.g.

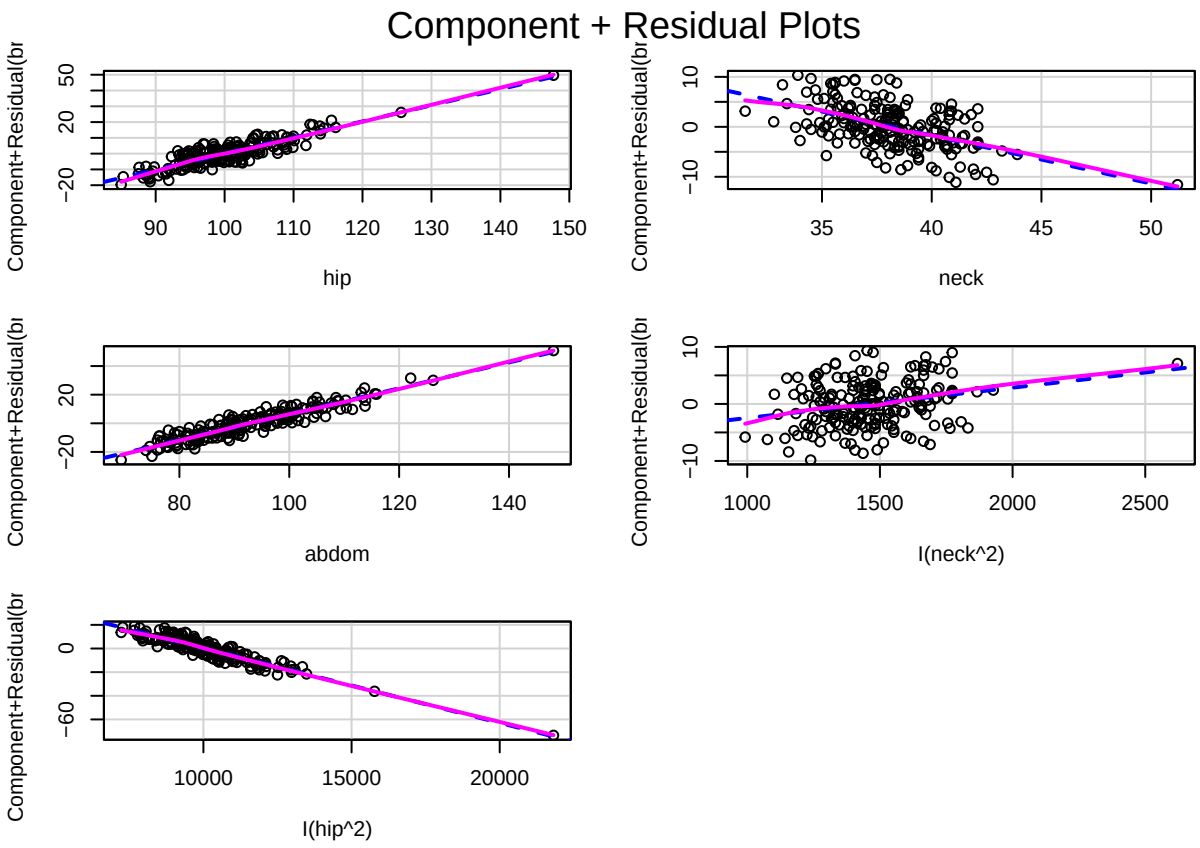
```
crPlots(fit.desired, terms=~hip+neck+abdom)
```




```
fit.fixed <- lm(brozek ~ abdom + neck + hip + I(hip^2) + I(neck^2) + wrist + forearm, data=Train)
residualPlots(fit.fixed, tests=FALSE)
```




```
crPlots(fit.fixed, terms=~hip+neck+abdom+I(neck^2)+I(hip^2))
```




```
summary(fit.fixed)
```

```
##
## Call:
## lm(formula = brozek ~ abdom + neck + hip + I(hip^2) + I(neck^2) +
##     wrist + forearm, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6388 -2.7809 -0.0926  2.7156  9.4176
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56.485394  41.153188  -1.373  0.17147
## abdom         0.904714   0.057088  15.848  < 2e-16 ***
## neck         -0.961824   3.467062  -0.277  0.78176
## hip           1.022190   0.787955   1.297  0.19608
## I(hip^2)     -0.006329   0.003787  -1.671  0.09628 .
## I(neck^2)     0.005322   0.045228   0.118  0.90645
## wrist        -1.314428   0.466348  -2.819  0.00532 **
## forearm       0.202652   0.188941   1.073  0.28480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.952 on 194 degrees of freedom
```

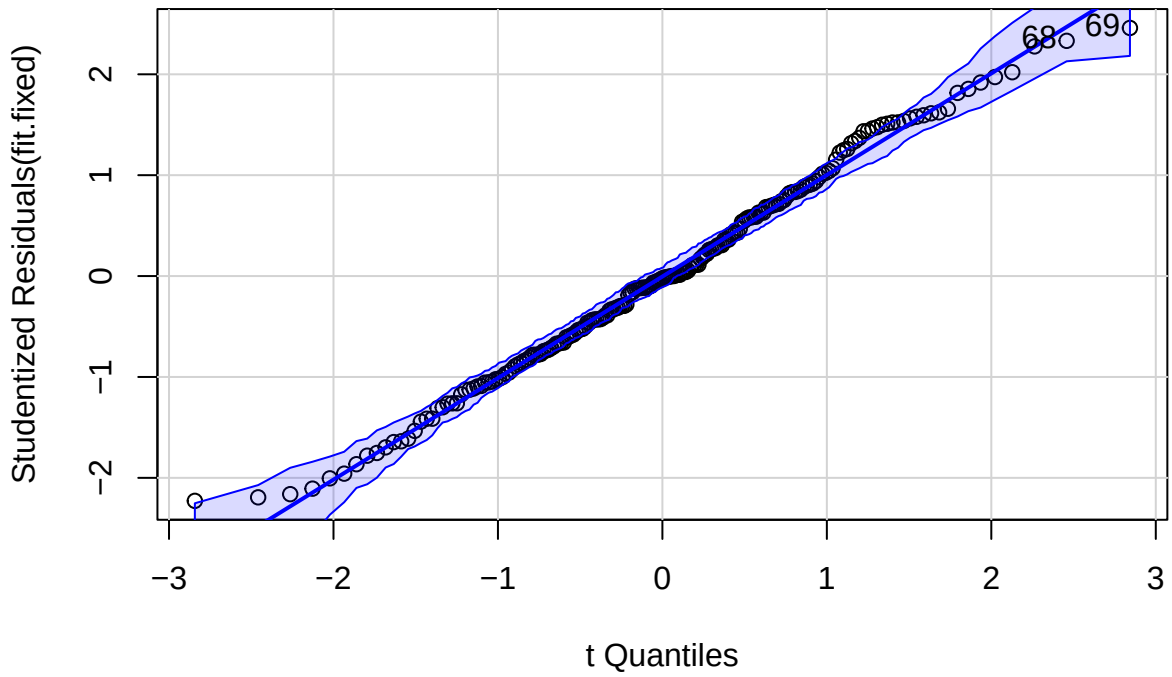
```
## Multiple R-squared:  0.7525, Adjusted R-squared:  0.7436
## F-statistic: 84.28 on 7 and 194 DF,  p-value: < 2.2e-16
```

After reconstructing the variables, the Component+Residual Plots look more compact in different variables than previously do. And the pearson residual plot connects all the variables together to be similar shape(it implies that our model regression are much more correct). Moreover, the p-value of such model is $< 2.2e\text{-}16$, thus we are more confident to reject the null hypothesis. Thus, our new model stands its logic.

From these graphs compared with the previous one, the fit.fixed model is a more procise one based on the assumptions of the residuals. Afterwards, we check the normality of this model.

QQplot

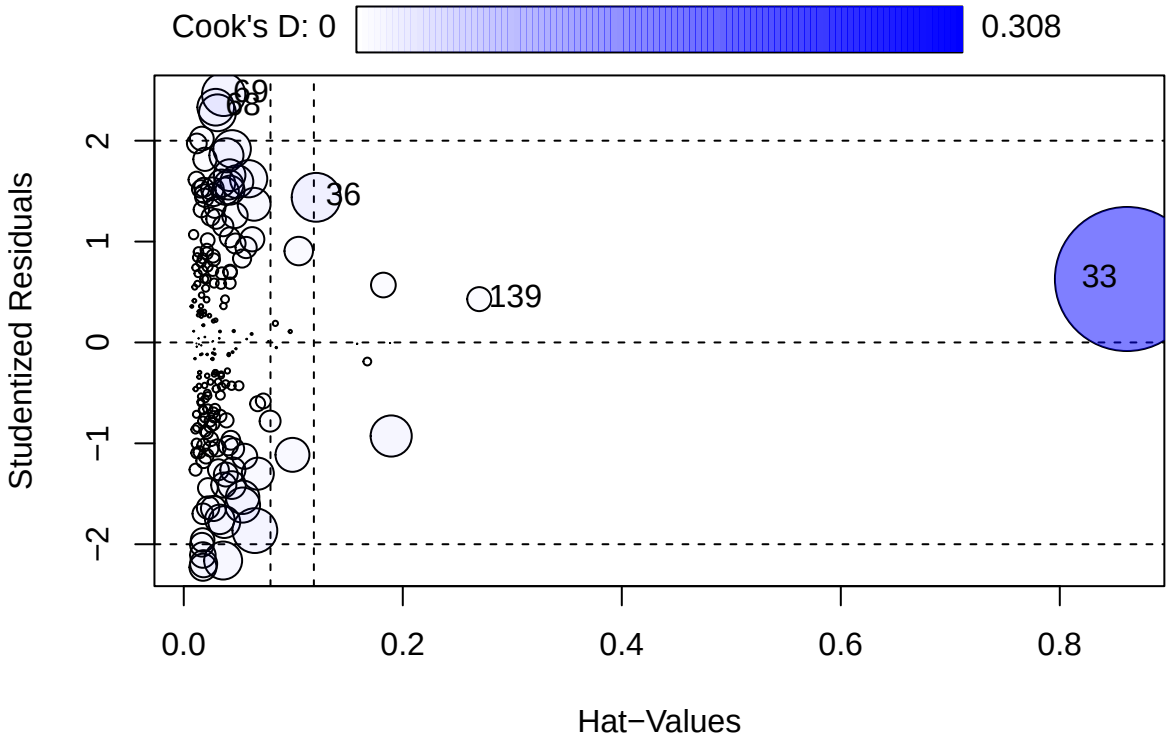
```
qqPlot(fit.fixed)
```




```
## [1] 68 69
```

This QQ plot looks excellent, the points mostly fall on or near the diagonal line. And there is no evidence to suggest that the residuals are not normally distributed.

```
influencePlot(fit.fixed)
```




```
##        StudRes        Hat       CookD
## 33   0.6288223 0.86131569 0.307933157
## 36   1.4384433 0.12087383 0.035366310
## 68   2.3317128 0.02919636 0.019981897
## 69   2.4592422 0.03643942 0.027864371
## 139 0.4294913 0.26981465 0.008556195
```

This model is approachable and acceptable since it has less value of Cook's distance, in fact none of the value of Cook's distance is larger than 1(the only proble is 0.307933157, slightly larger than other values, but we ignore this since its subtlety), and the qqplot of it looks fine. All in all, we decide to take this model (fit.fixed) as our final configuration.

What are left to do is to run the data in the test file and examine our model fit.fixed to be a rather powerful one or not.

Part(b)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Test <- read.table(file='Test.txt', header=TRUE,sep="")
Test[1:10,]
```

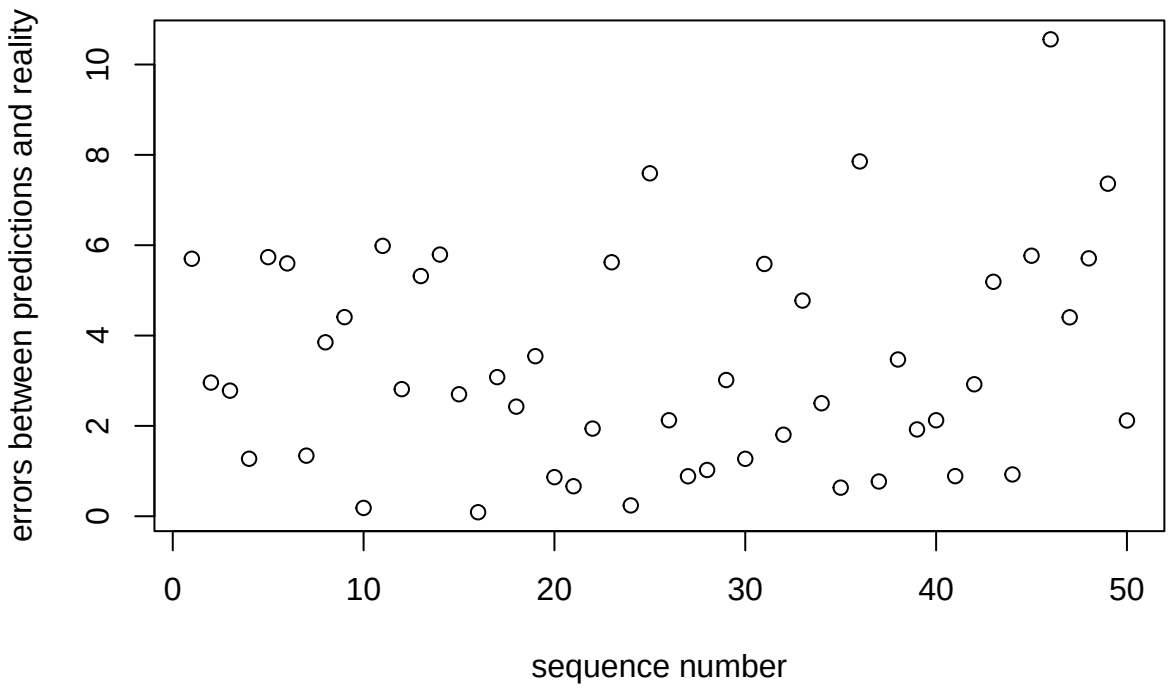
```
##    brozek neck chest abdom   hip thigh knee ankle biceps forearm wrist
## 1    24.6 34.0  95.8  87.9  99.2  59.6 38.9  24.0   28.8    25.2  16.6
## 2    20.6 39.0 104.5  94.4 107.8  66.0 42.0  25.6   35.7    30.6  18.8
## 3    12.8 37.8  99.6  88.5  97.1  60.0 39.4  23.2   30.5    29.0  18.8
## 4    20.5 36.4  99.1  92.8  99.2  63.1 38.7  21.7   31.1    26.4  16.9
## 5    28.1 38.9 101.9  96.4 105.2  64.8 40.8  23.1   36.2    30.8  17.3
## 6     6.5 37.3  93.5  84.5 100.6  58.5 38.8  21.5   30.1    26.4  17.9
## 7    30.4 37.3 113.3 111.2 114.1  67.7 40.9  25.0   36.7    29.8  18.4
## 8    11.2 33.6  88.2  73.7  88.5  53.3 34.5  22.5   27.9    26.2  17.3
## 9     6.4 34.6  89.8  79.5  92.7  52.7 37.5  21.9   28.8    26.8  17.9
## 10    5.0 34.0  83.4  70.4  87.2  50.6 34.4  21.9   26.8    25.8  16.8
```

```
# We select the true value of brozek in the Test file.
TestResponses = select(Test, brozek)$brozek
TestResponses
```

```
##  [1] 24.6 20.6 12.8 20.5 28.1  6.5 30.4 11.2  6.4  5.0 10.7  9.1 18.6 26.2 22.6
## [16] 14.3 11.1 19.8 21.9 17.8 21.0 20.9 25.9 16.7 25.1 14.8 20.4 26.3 18.1 10.8
## [31] 28.6 19.1 10.6 30.1 19.0  4.1 25.8 11.9 20.2 23.8 11.8 24.0 21.5  7.3 21.7
## [46] 31.7 10.1 12.7 11.1 33.6
```

We then use our model to make predictions and compared with the real ones stored in the TestResponses. Apart from that, we try to use a powerful tool named expected square prediction error(MSE), which can imply the powerness of our model if the value is relatively small.

```
predictions <- predict(fit.fixed, newdata=select(Test, -brozek))
errors <- abs(predictions - TestResponses)
plot(errors, xlab='sequence number', ylab='errors between predictions and reality')
```




```
mse_MSE <- mean((predictions - TestResponses)^2)
mse_MSE
```

```
## [1] 16.8382
```

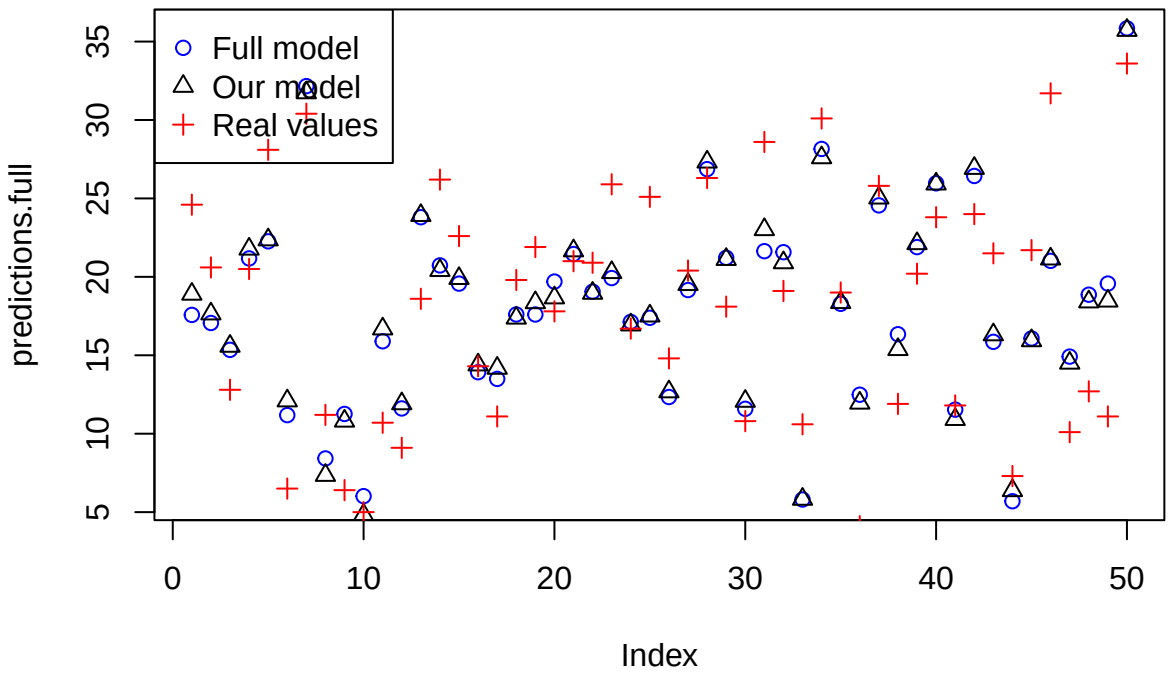
How does this prediction perform compared with a full model(fit as previous defined): brozek = intercept + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist

```
predictions.full <- predict(fit, newdata=select(Test, -brozek))
mse_MSE.full <- mean((predictions.full - TestResponses)^2)
mse_MSE.full
```

```
## [1] 18.299
```

We then combine all the predictions and real values in one graph, and compare them statistically.

```
plot(predictions.full, col='blue')
points(predictions, col = 'black', pch=2)
points(TestResponses, col = 'red', pch=3)
legend("topleft", legend = c("Full model", "Our model", "Real values"), pch = 1:3, col = c("blue", "bla
```



Clearly, our model returns a more concise value compared to the full model in the question.

The mse obtained is 18.299 which is larger than the previous model(fit.desired) of that 16.8382. In this case, the fit.desired model (what our final model is) is significantly more powerful than the full model (fit).

Besides, all the summaries data or other plots are combined in the previous page. Thus, after all the analysis, our ultimate best model is set to be: brozek ~ abdom + neck + hip + I(hip^2) + I(neck^2) + wrist + forearm, which is a linear model.