Robust Statistics Assignment

Janne Flinck

79428

1

I would use robust statistical methods because they are more resistant to small deviations from the assumptions, in other words robust with respect to outliers and stable with respect to small deviations from the assumed parametric model. It is well known that classical methods behave quite poorly under slight violations of the strict model assumptions. In particular, least squares estimate for regression models are highly sensitive to outliers. If the outlier results from non-normal measurement error or some other violation of standard ordinary least squares assumptions, then it compromises the validity of the regression results if a non-robust regression technique is used. Another instance in which robust estimates should be used is when there is a suspicion of heteroscedasticity.

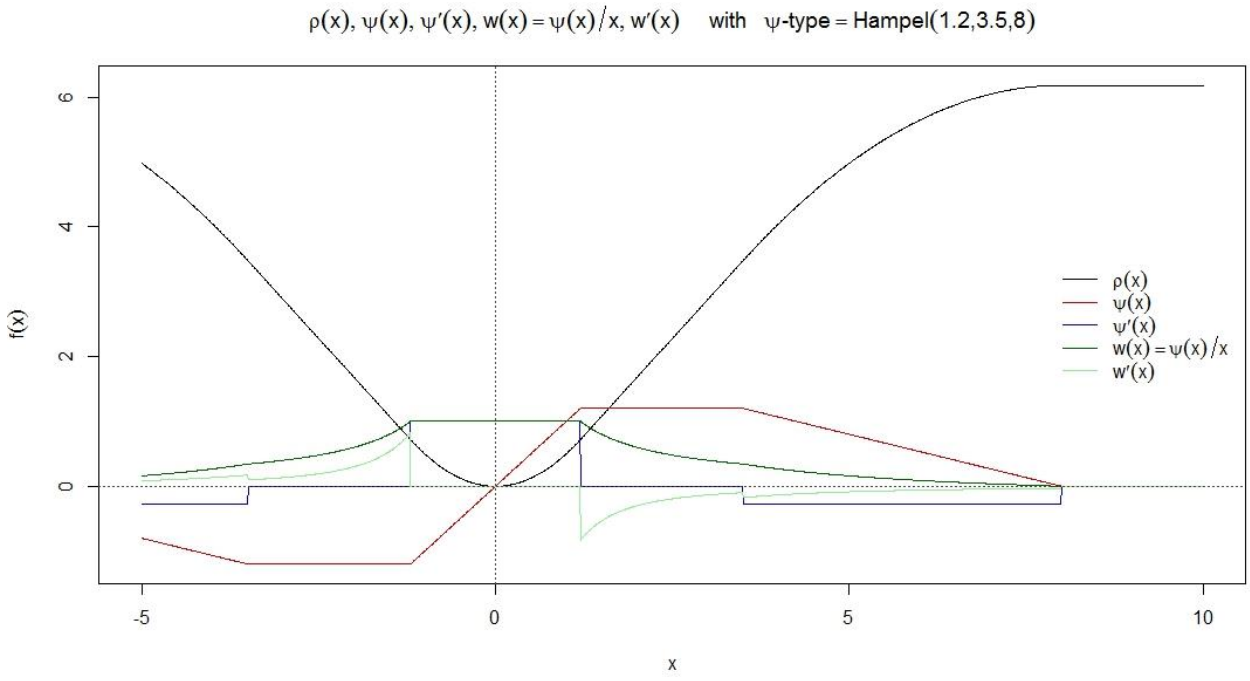
It is discouraged to screen the data, to remove outliers and then to apply classical inferential methods. This is not a simple and good way to proceed because those outliers might be valid regarding the data generation process. In multivariate or highly structured data, it can be hard to single out outliers or to identify influential points. In place of rejecting an observation, it could be better to down-weight uncertain observations. Rejecting outliers reduces the sample size, could affect the distribution theory, and variances could be underestimated from the cleaned data.

I would use robust methods to compliment and to compare to the classical methods. In other words, I would always make at least two models, one classical and one robust, and see which one fit the data better. This is especially true if I suspect that the data has outliers or influence points. If they give about the same results, I would use the classical method because it is easier to explain to someone not familiar with statistics. However, if the robust method gives a more accurate result, I would definitely use it.

2

I get the following location estimate

```
4.58924
```



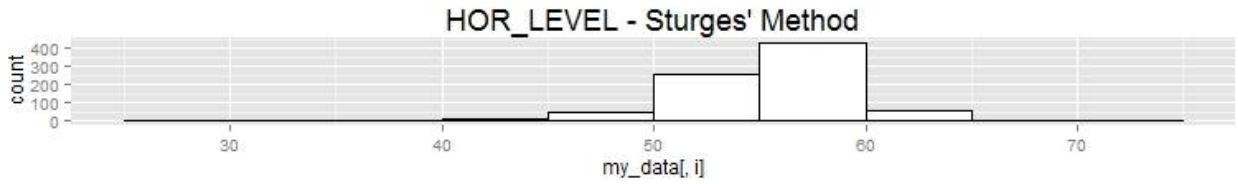
It is a redescender.

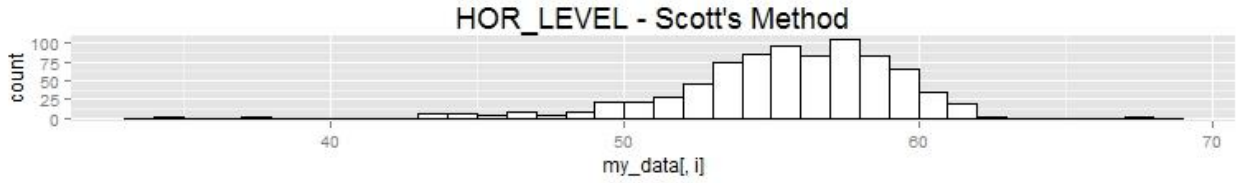
3

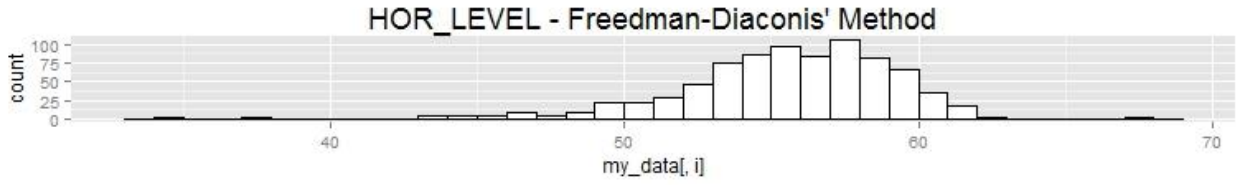
4

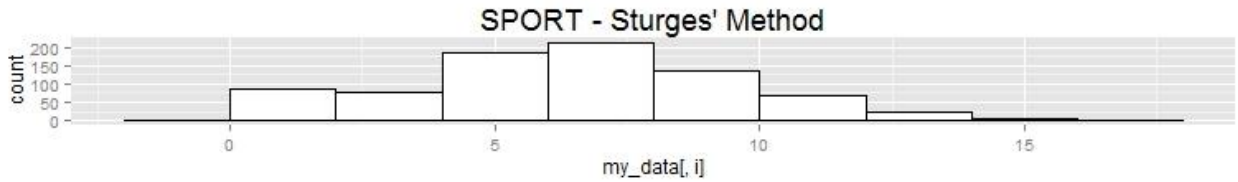
Data Understanding

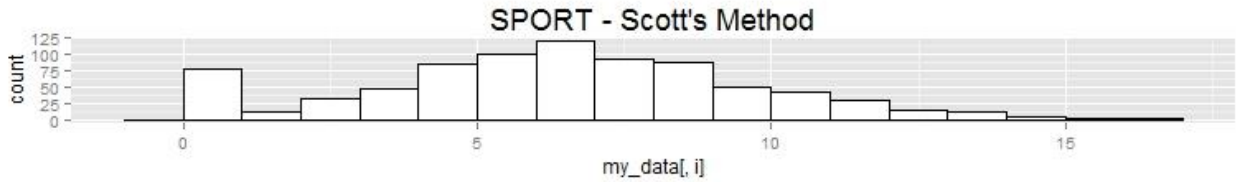
I start off by plotting histograms of the attributes with a different number of bins.

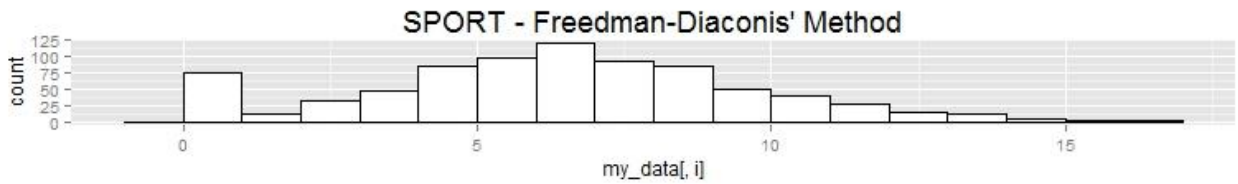


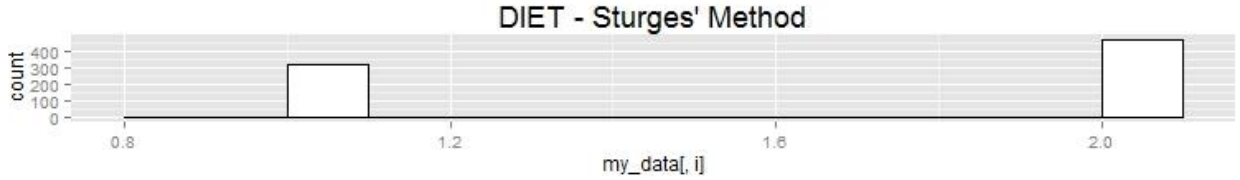


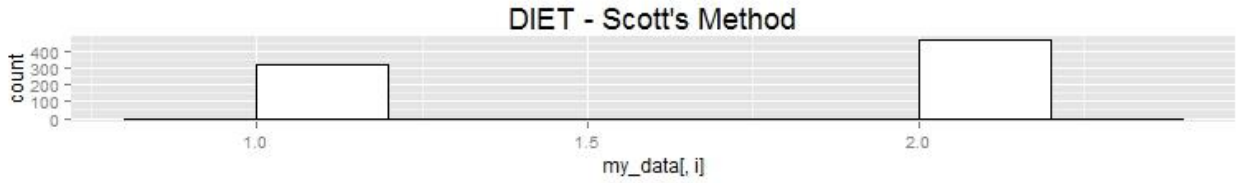


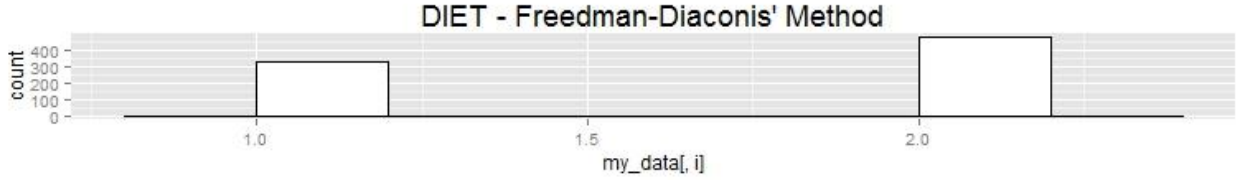


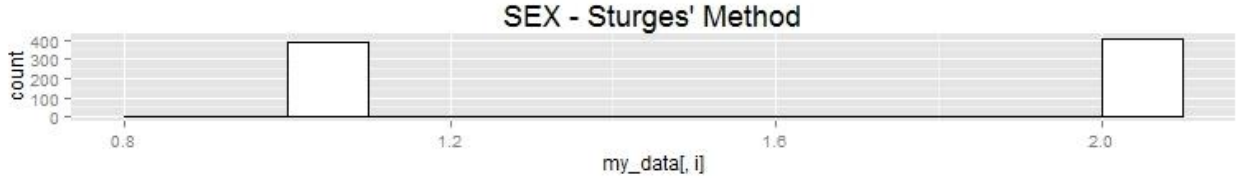


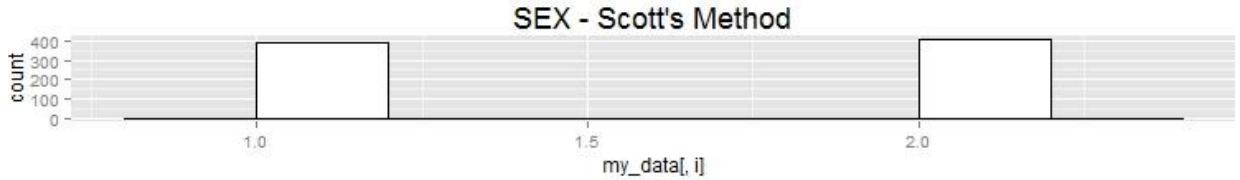


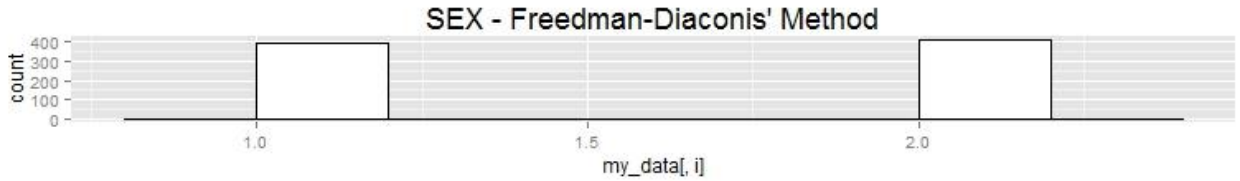

DIET - Sturges' Method
count
400
300
200
100
0
0.8
1.2
1.6
2.0
my_data[, i]

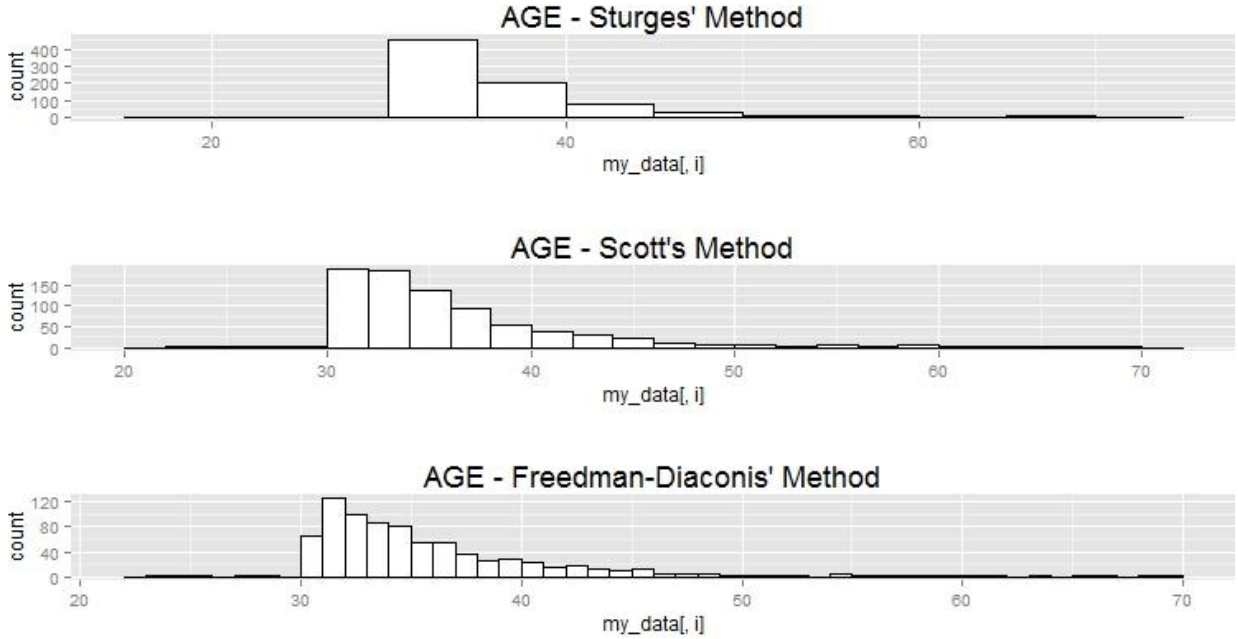

DIET - Scott's Method
count
400
300
200
100
0
1.0
1.5
2.0
my_data[, i]


DIET - Freedman-Diaconis' Method
count
400
300
200
100
0
1.0
1.5
2.0
my_data[, i]


SEX - Sturges' Method
count
400
300
200
100
0
0.8
1.2
1.6
2.0
my_data[, i]


SEX - Scott's Method
count
400
300
200
100
0
1.0
1.5
2.0
my_data[, i]

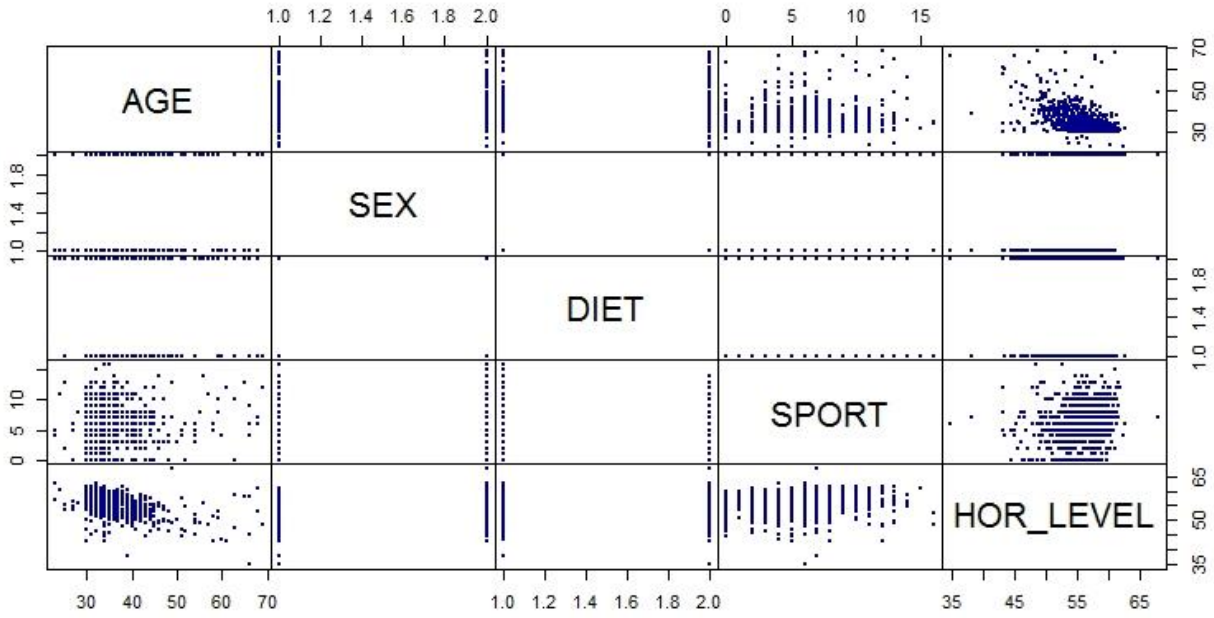

SEX - Freedman-Diaconis' Method
count
400
300
200
100
0
1.0
1.5
2.0
my_data[, i]

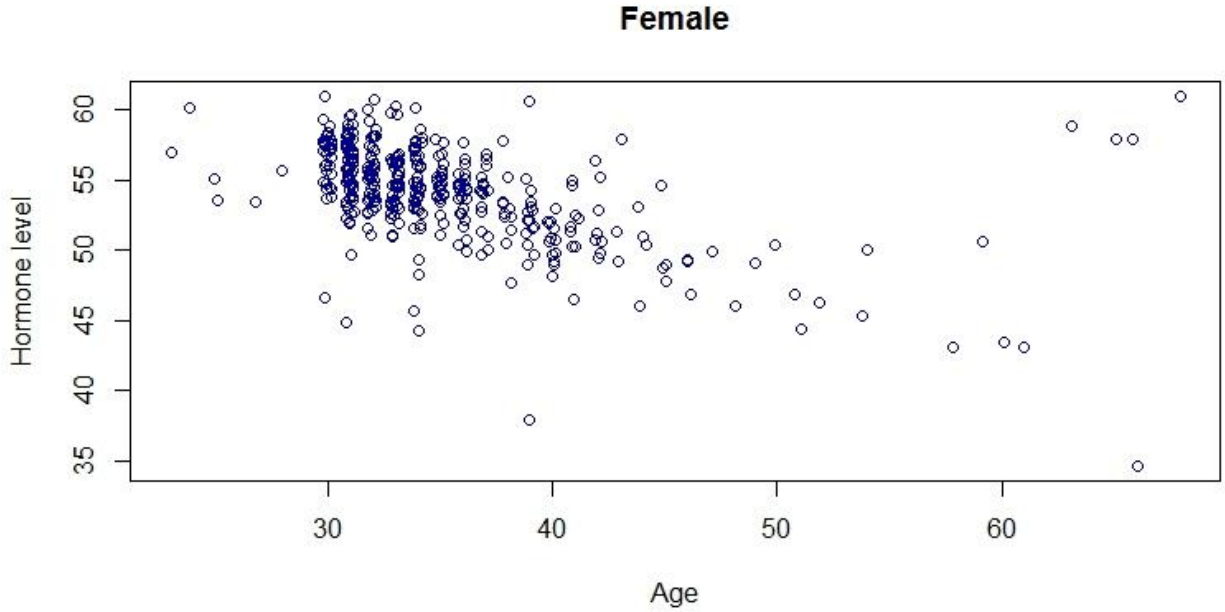


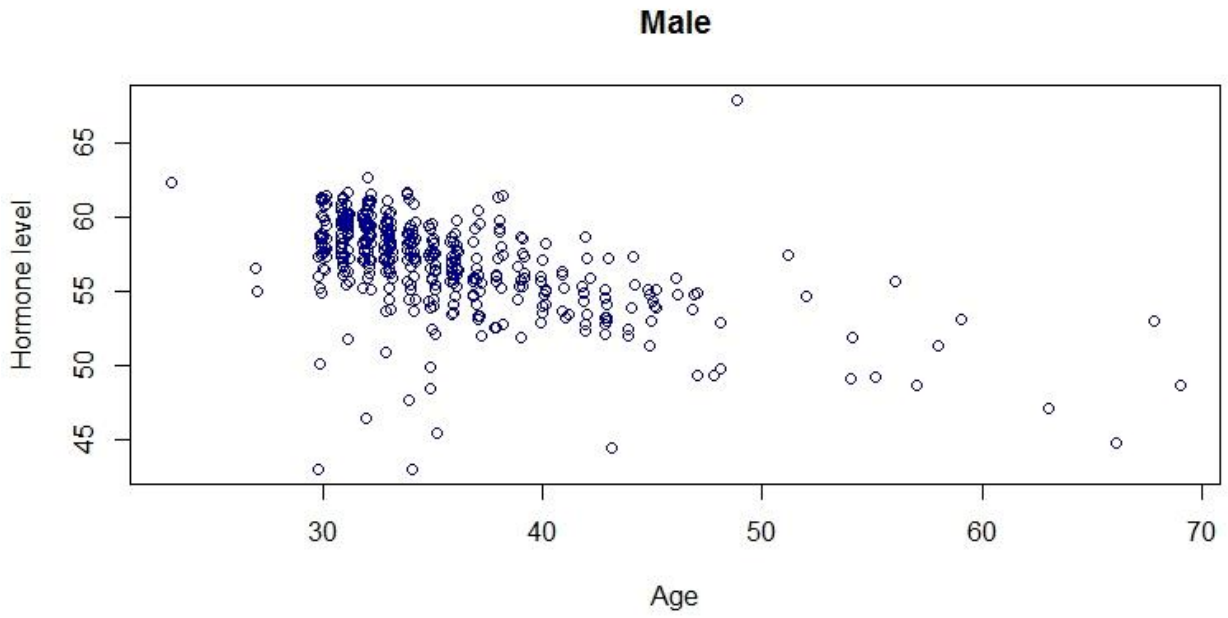
From these histograms we can observe the following things:

- HOR_LEVEL has a negative skewness and AGE has a positive skewness and a high kurtosis
- HOR_LEVEL and AGE have outliers.
- SPORT could be considered a normal distribution, if it did not have a spike at value zero.
- There are more people with Unhealthy diets.
- There seem to be about an equal amount of men and women.
- The continuous variables have vastly different means and variances.

I have produced a scatter plot of the data. We can see that AGE and HOR_LEVEL are somewhat negatively correlated.

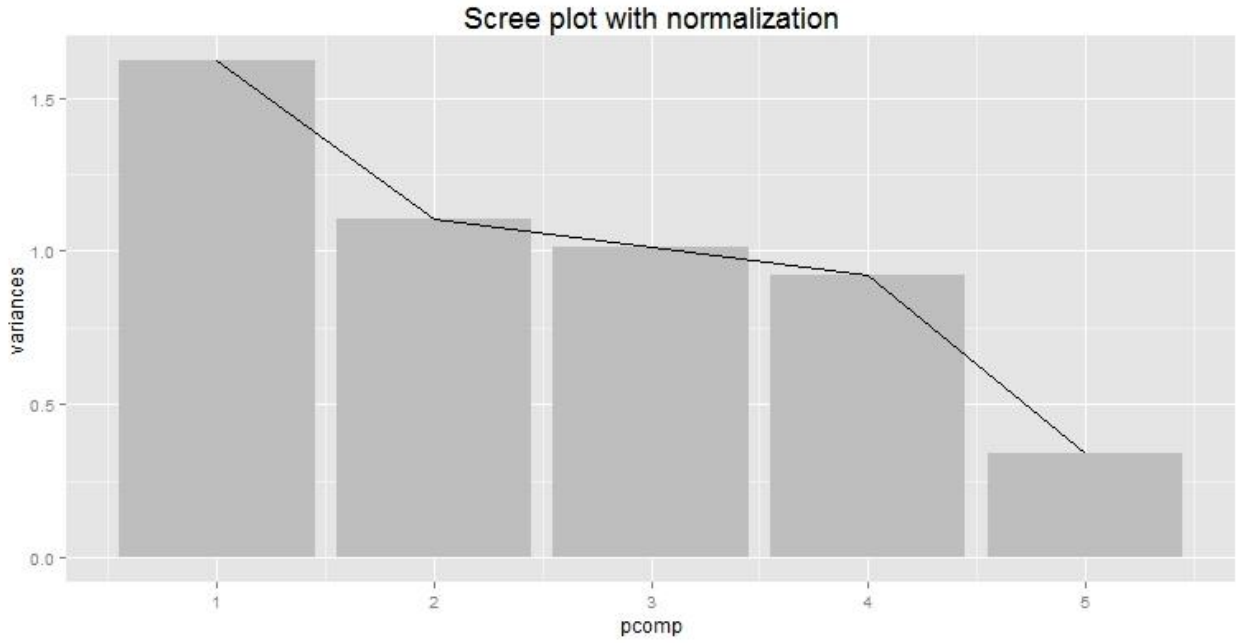


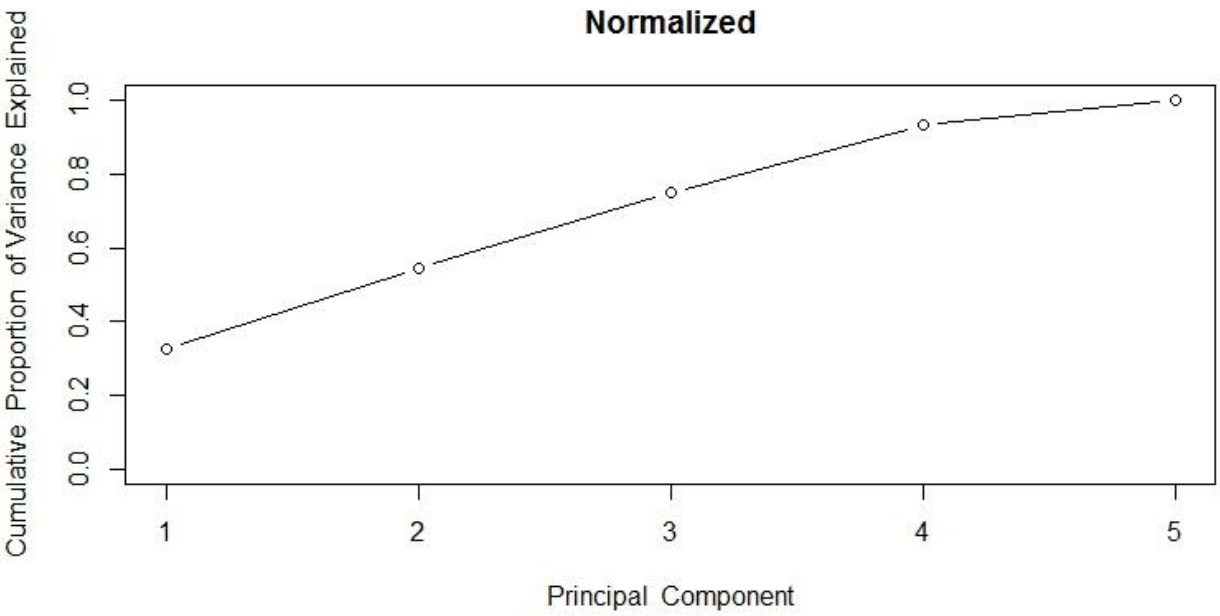




HOR_LEVEL and AGE look negatively correlated for both men and women separately; however there are some interesting values for females with a high age.

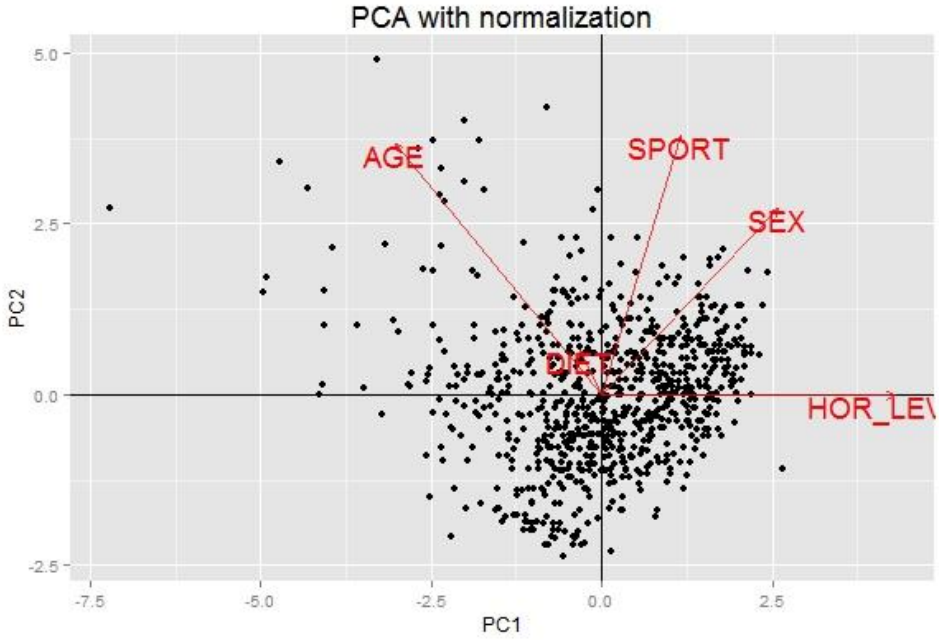
Here are the scree plots of the data:





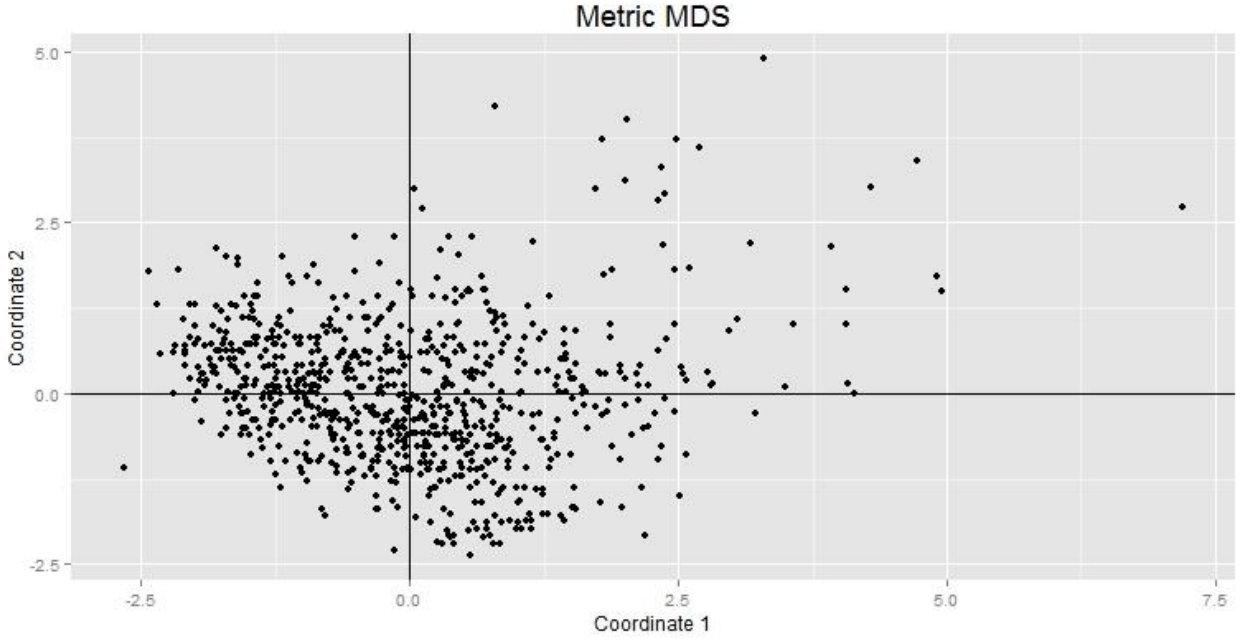
From these plots we can see that the first principle component explains about 35% of the variance and the second principle explains about 20% of the variance. Together they explain about 55%. A scree plot allows a graphical assessment of the relative contribution of the PCs in explaining the variability of the data.

Here's the plot with PC1 on the horizontal axis and PC2 on the vertical axis:



What's interesting about this plot is that judging by the first two principal components; a HOR_LEVEL doesn't seem to be strongly correlated with any variable. In the above plot we can see some points clearly being away from the center of the point cloud. These appear to be in the top left quadrant.

Next I have plotted the 2D MDS chart. Instead of preserving the variance, like in PCA, MDS preserves the Euclidean distance in the data. For this data it looks exactly the same as PCA, we can see some clear outliers.



Correlations:

Kendall's Tau

	AGE	SPORT	HOR_LEVEL
AGE	1.000	0.035	-0.342
SPORT	0.035	1.000	0.172
HOR_LEVEL	-0.342	0.172	1.000

Pearson's Correlation

	AGE	SPORT	HOR_LEVEL
AGE	1.000	0.078	-0.474
SPORT	0.078	1.000	0.194
HOR_LEVEL	-0.474	0.194	1.000

Highest absolute correlation is between HOR_LEVEL and AGE, just like was assumed from the visual inspection of the data.

Data Modeling

Next I did some basic models of the data.

```
Call:
lm(formula = HOR_LEVEL ~ . - ID, data = DATA.ANA)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7550  -1.1614   0.1098   1.4765  16.1786

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.24287    0.59176 106.873  < 2e-16 ***
AGE           -0.29087    0.01567 -18.567  < 2e-16 ***
SEXmale        2.75557    0.19213  14.342  < 2e-16 ***
DIETunhealthy -0.33621    0.19526  -1.722   0.0855 .
SPORT          0.23886    0.03006   7.946  6.6e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.712 on 795 degrees of freedom
Multiple R-squared:  0.4273,   Adjusted R-squared:  0.4245
F-statistic: 148.3 on 4 and 795 DF,  p-value: < 2.2e-16
```

In this basic model we can see that all variables are statistically significant at the 0.05 level and the model has an adjusted R-squared of 0.42. Age has a negative effect on hormonal level, being male has a higher positive effect on hormonal level than for females, an unhealthy diet has a negative effect on hormonal level when compared to a healthy diet and sport has a positive effect on hormonal level.

Here are the confidence intervals for the predictors:

```
                   2.5 %      97.5 %
(Intercept)   62.0812749 64.40445862
AGE           -0.3216241 -0.26011901
SEXmale        2.3784303  3.13270530
DIETunhealthy -0.7194907  0.04707989
SPORT          0.1798514  0.29787304
```

DIETunhealthy is the only one that contains 0 in the confidence interval.

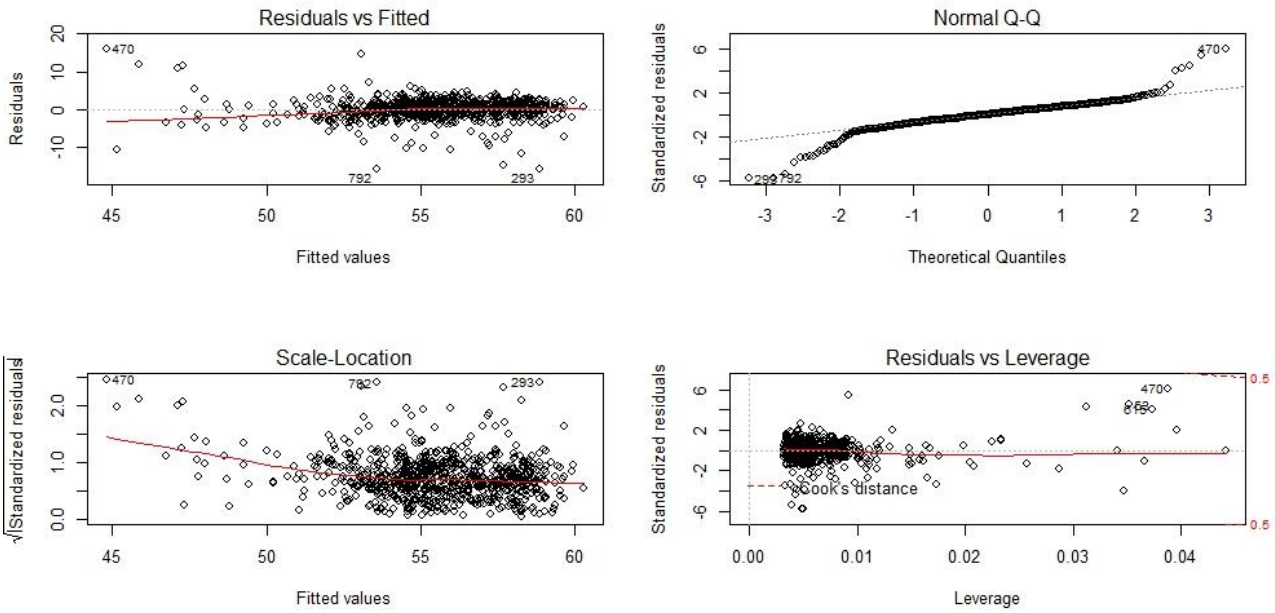
Now that I have a model, I will look for OLS assumption violations for continuous variables (HOR_LEVEL and AGE).

First I check if there a relation between the ith and the i+1th observation, with the following results. (The lag might be larger, i.e. the ith and i+kth observation might be correlated).

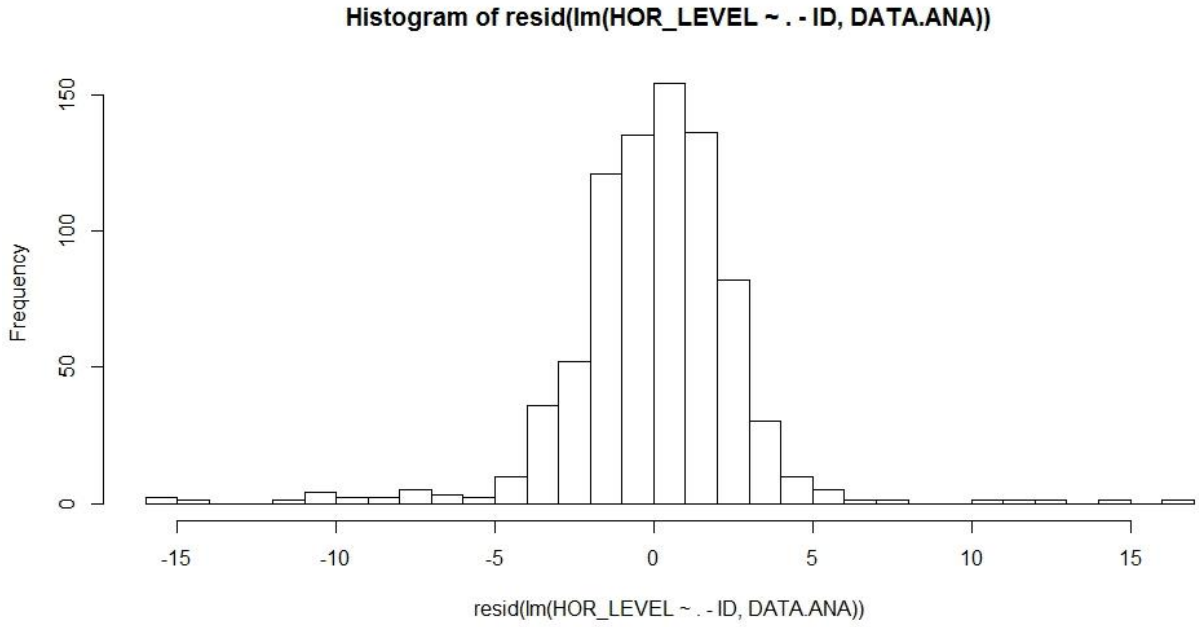
```
> cor(ylag,ynew)
[1] -0.01851628
> cor(cnew,clag)
[1] 0.01502376
```

No correlation, so I assume that the data is generated by an i.i.d process.

Do residuals and yhats correlate? Does explanatory variable correlate with the estimated residuals? Is the relationship linear?



There seems to be no relation between residuals and fitted values so Cov(X'u)=0 is fine. From the QQ-plot we see some departures from the line, which brings doubts about a normal distribution. There are also influential points in the data.



The residuals do not follow a normal distribution, as can be seen from this histogram. Just to make sure I perform a shapiro-wilk test of normality:

```
        Shapiro-Wilk normality test
data:  resid(lm(HOR_LEVEL ~ . - ID, DATA.ANA))
W = 0.8846, p-value < 2.2e-16
```

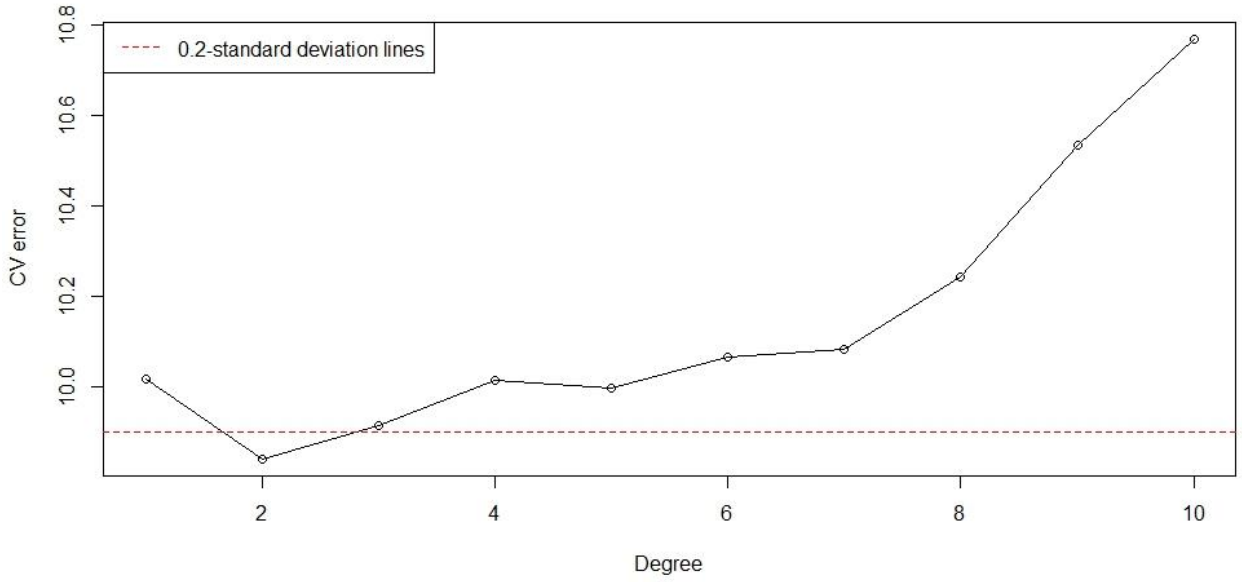
H0 (normality) is rejected.

I will still try to find the best lm model for the data. I will add interaction terms and check for polynomial degrees.

```
Call:
lm(formula = HOR_LEVEL ~ (AGE + SEX + DIET + SPORT)^2, data = DATA.ANA)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6216  -1.1708   0.1693   1.4781  15.9421

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           68.796722   1.529067  44.993  < 2e-16 ***
AGE                   -0.435247   0.042798 -10.170  < 2e-16 ***
SEXmale                1.920351   1.173684   1.636 0.102202
DIETunhealthy         -2.499084   1.183330  -2.112 0.035008 *
SPORT                 -0.426444   0.164201  -2.597 0.009577 **
AGE:SEXmale            0.024985   0.031321   0.798 0.425282
AGE:DIETunhealthy      0.035321   0.031791   1.111 0.266893
AGE:SPORT              0.016670   0.004463   3.735 0.000201 ***
SEXmale:DIETunhealthy  0.156473   0.388322   0.403 0.687097
SEXmale:SPORT         -0.015732   0.060185  -0.261 0.793863
DIETunhealthy:SPORT    0.140494   0.060647   2.317 0.020781 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.686 on 789 degrees of freedom
Multiple R-squared:  0.4426,   Adjusted R-squared:  0.4355
F-statistic: 62.65 on 10 and 789 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Model 1: HOR_LEVEL ~ (ID + AGE + SEX + DIET + SPORT) - ID
Model 2: HOR_LEVEL ~ (AGE + SEX + DIET + SPORT)^2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    795 5846.8
2    789 5691.2  6    155.58 3.5947 0.001598 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model with interaction terms is better than the basic model.

To see, which polynomial would be best for HOR_LEVEL and AGE, I use 10-fold cross validation. Here are the error results for the polynomial degrees:



As can be seen from the plot, a polynomial degree of two minimizes the error.

Now I can fit a model with interaction terms and AGE with a polynomial.

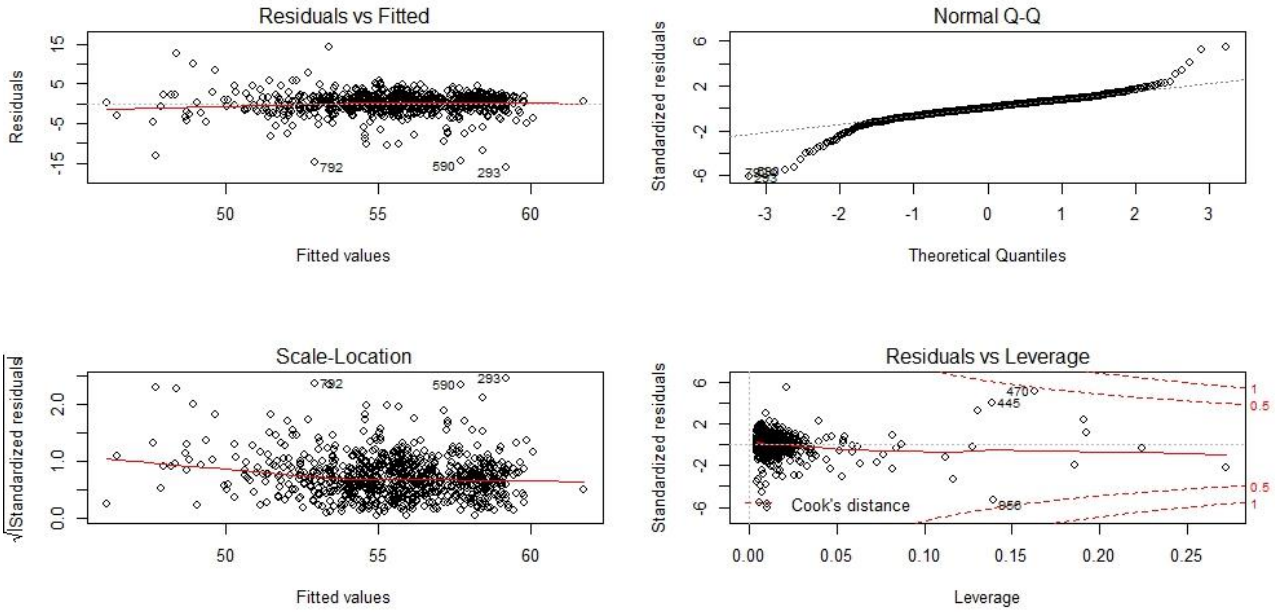
```
Call:
lm(formula = HOR_LEVEL ~ . - ID + I(AGE^2) + (AGE + SEX + DIET +
    SPORT)^2, data = DATA.ANA)

Residuals:
     Min       1Q   Median       3Q      Max
-16.1120  -1.1083   0.1356   1.4726  14.4412

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           78.534420   2.521653  31.144  < 2e-16 ***
AGE                   -0.932688   0.111539  -8.362 2.78e-16 ***
SEXmale                1.739813   1.158109   1.502   0.1334
DIETunhealthy         -2.850328   1.169290  -2.438   0.0150 *
SPORT                 -0.251297   0.165968  -1.514   0.1304
I(AGE^2)               0.006136   0.001274   4.818 1.74e-06 ***
AGE:SEXmale            0.028536   0.030898   0.924   0.3560
AGE:DIETunhealthy      0.043983   0.031404   1.401   0.1617
AGE:SPORT              0.011568   0.004527   2.555   0.0108 *
SEXmale:DIETunhealthy  0.256923   0.383535   0.670   0.5031
SEXmale:SPORT         -0.011794   0.059361  -0.199   0.8426
DIETunhealthy:SPORT    0.141793   0.059811   2.371   0.0180 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.649 on 788 degrees of freedom
Multiple R-squared:  0.4585,   Adjusted R-squared:  0.451
F-statistic: 60.66 on 11 and 788 DF,  p-value: < 2.2e-16
```

It is interesting to see that AGE and SEXmale as such a high p-value in this model. There is a marginal increase in the adjusted R-squared compared to the previous model.

Confidence intervals:

```
                          coef  2.5 % 97.5 %
(Intercept)             78.534 73.584 83.484
AGE                     -0.933 -1.152 -0.714
SEXmale                  1.740 -0.534  4.013
DIETunhealthy           -2.850 -5.146 -0.555
SPORT                   -0.251 -0.577  0.074
I(AGE^2)                 0.006  0.004  0.009
AGE:SEXmale              0.029 -0.032  0.089
AGE:DIETunhealthy        0.044 -0.018  0.106
AGE:SPORT                0.012  0.003  0.020
SEXmale:DIETunhealthy    0.257 -0.496  1.010
SEXmale:SPORT           -0.012 -0.128  0.105
DIETunhealthy:SPORT      0.142  0.024  0.259
```

Residual plots:



Since not all variables were statistically significant I will test a model that does not have those variables, just to see if there is a significant difference.

```
Call:
lm(formula = HOR_LEVEL ~ . - ID + I(AGE^2) + AGE:SPORT + DIET:SPORT,
    data = DATA.ANA)
Residuals:
     Min       1Q   Median       3Q      Max
-16.2447  -1.1119   0.1768   1.4746  14.9023
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         76.688997   2.279711  33.640  < 2e-16 ***
AGE                 -0.874716   0.106399  -8.221 8.22e-16 ***
SEXmale              2.835545   0.188039  15.080  < 2e-16 ***
DIETunhealthy       -1.192893   0.397426  -3.002  0.00277 **
SPORT               -0.247191   0.164868  -1.499  0.13419
I(AGE^2)             0.005967   0.001268   4.704 3.00e-06 ***
AGE:SPORT            0.011150   0.004487   2.485  0.01315 *
DIETunhealthy:SPORT  0.147169   0.059359   2.479  0.01337 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 2.647 on 792 degrees of freedom
Multiple R-squared:  0.4564,   Adjusted R-squared:  0.4516
F-statistic: 94.98 on 7 and 792 DF,  p-value: < 2.2e-16

Confidence intervals:
                      coef  2.5 % 97.5 %
(Intercept)         76.689 72.214 81.164
AGE                 -0.875 -1.084 -0.666
SEXmale              2.836  2.466  3.205
DIETunhealthy       -1.193 -1.973 -0.413
SPORT               -0.247 -0.571  0.076
I(AGE^2)             0.006  0.003  0.008
AGE:SPORT            0.011  0.002  0.020
DIETunhealthy:SPORT  0.147  0.031  0.264
```

ANOVA:

```
Analysis of Variance Table

Model 1: HOR_LEVEL ~ (ID + AGE + SEX + DIET + SPORT) - ID
Model 2: HOR_LEVEL ~ (AGE + SEX + DIET + SPORT)^2
Model 3: HOR_LEVEL ~ (ID + AGE + SEX + DIET + SPORT) - ID + I(AGE^2) +
    (AGE + SEX + DIET + SPORT)^2
Model 4: HOR_LEVEL ~ (ID + AGE + SEX + DIET + SPORT) - ID + I(AGE^2) +
    AGE:SPORT + DIET:SPORT
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1    795 5846.8
2    789 5691.2  6   155.578  3.6960  0.001249 **
3    788 5528.4  1   162.862 23.2141 1.739e-06 ***
4    792 5550.6 -4   -22.228  0.7921  0.530451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

AICS:

```
> AIC(fit)
[1] 3873.532
> AIC(fit.1)
[1] 3863.956
> AIC(fit.2)
[1] 3842.729
> AIC(fit.3)
[1] 3837.939
```

The lower the AIC, the better. Fit.3 seems to be the best one here.

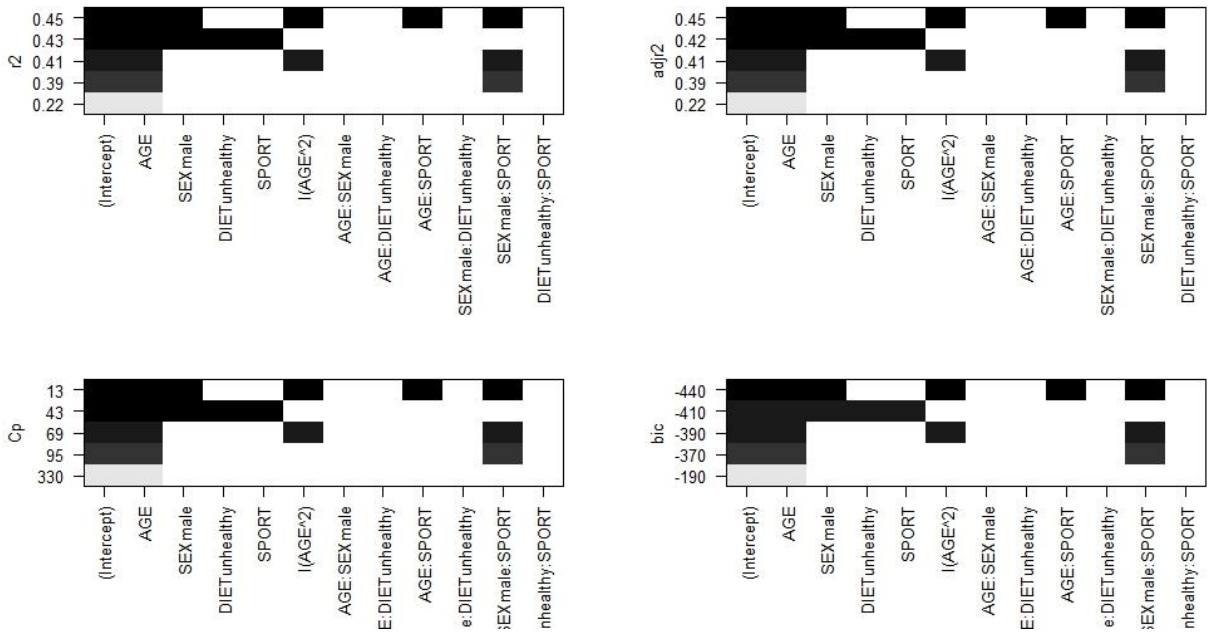
Influence points:

```
      cov.r   cook.d hat
3    0.91_*  0.01   0.01
53   0.98    0.14   0.13_*
54   1.05_*  0.00   0.03
55   1.05_*  0.00   0.04
68   1.09_*  0.01   0.08_*
76   0.93_*  0.04   0.05_*
79   0.92_*  0.02   0.03
80   0.98    0.02   0.05_*
133  0.86_*  0.01   0.01
192  0.81_*  0.01   0.01
200  1.12_*  0.02   0.11_*
229  0.84_*  0.00   0.00
231  0.99    0.01   0.03
293  0.57_*  0.03   0.01
308  0.90_*  0.01   0.02
344  0.95_*  0.01   0.01
348  1.08_*  0.01   0.08_*
```

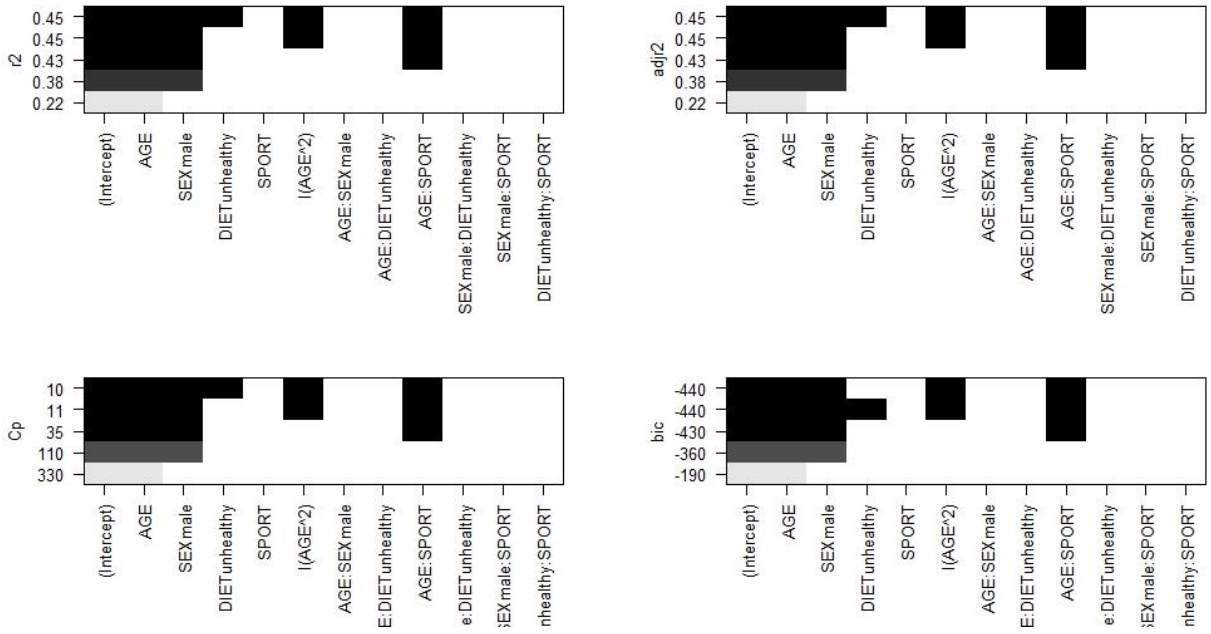
349	1.16_*	0.00	0.13_*
374	1.06_*	0.00	0.04
382	1.06_*	0.00	0.05_*
387	0.97	0.12	0.12_*
418	0.98	0.02	0.04
433	1.05_*	0.02	0.07_*
445	0.92_*	0.22	0.14_*
470	0.80_*	0.44	0.16_*
479	1.11_*	0.00	0.08_*
483	0.94_*	0.00	0.01
486	0.92_*	0.02	0.03
492	0.65_*	0.06	0.02
494	1.30_*	0.15	0.27_*
529	1.07_*	0.00	0.05_*
539	0.81_*	0.01	0.01
567	1.05_*	0.00	0.04
582	1.02	0.04	0.08_*
590	0.64_*	0.02	0.01
591	1.05_*	0.00	0.03
615	1.15_*	0.11	0.19_*
627	1.05_*	0.00	0.03
629	0.94_*	0.01	0.02
632	1.05_*	0.00	0.03
638	1.07_*	0.00	0.06_*
647	1.07_*	0.00	0.05_*
673	1.11_*	0.00	0.09_*
693	0.90_*	0.01	0.02
704	1.05_*	0.00	0.05_*
713	1.31_*	0.00	0.22_*
720	1.08_*	0.00	0.06_*
753	0.95_*	0.01	0.02
760	0.74_*	0.01	0.01
764	1.18_*	0.07	0.19_*
773	1.23_*	0.03	0.19_*
774	1.05_*	0.00	0.03
792	0.62_*	0.03	0.01
804	0.80_*	0.01	0.01
830	0.89_*	0.01	0.01
836	1.07_*	0.00	0.05_*
840	1.03	0.02	0.06_*
844	1.05_*	0.00	0.03
928	1.08_*	0.00	0.06_*
956	0.76_*	0.38	0.14_*

Next I will do variable selection based on R-squared, Adjusted R-squared, Mallow's Cp and BIC. Then I will create a robust model.

Forward stepwise selection:



Backward stepwise selection:

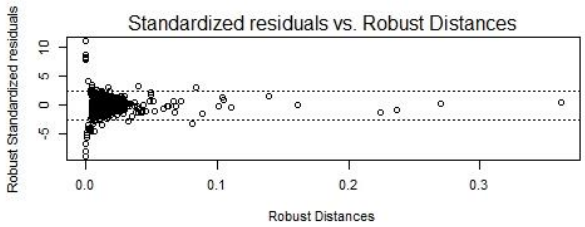


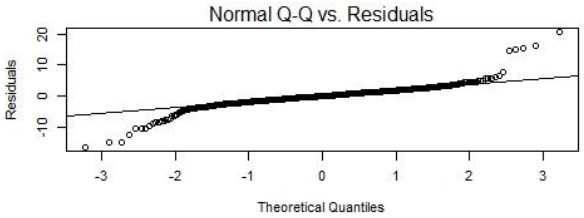
It is interesting to see that all selection criterion would drop the AGE:SEX variable. I will still run a model with all the variables, their synergetic effects and AGE to the power of 2. I chose the setting KS2014 because it is the one suggested to be state of the art in the lecture slides and in the package description of lmrob.

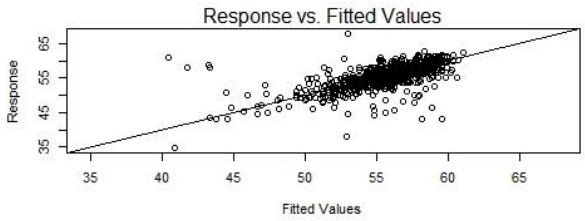
```
Call:
lmrob(formula = HOR_LEVEL ~ . - ID + I(AGE^2) + (AGE + SEX + DIET + SPORT)^2,
    data = DATA.ANA, method = "MM", setting = "KS2014")
 \--> method = "MM"
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           71.2885500  2.0267760  35.173  < 2e-16 ***
AGE                   -0.5544452  0.0930716  -5.957 3.87e-09 ***
SEXmale               -1.3500132  0.9396627  -1.437   0.1512
DIETunhealthy         -1.1696499  0.9230052  -1.267   0.2055
SPORT                  0.1789987  0.1274097   1.405   0.1604
I(AGE^2)               0.0012630  0.0011009   1.147   0.2516
AGE:SEXmale            0.1181428  0.0254573   4.641 4.06e-06 ***
AGE:DIETunhealthy      0.0014755  0.0252359   0.058   0.9534
AGE:SPORT              0.0001555  0.0035004   0.044   0.9646
SEXmale:DIETunhealthy  0.5117789  0.2797045   1.830   0.0677 .
SEXmale:SPORT         -0.0101384  0.0437987  -0.231   0.8170
DIETunhealthy:SPORT    0.1009363  0.0440880   2.289   0.0223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 1.869
Multiple R-squared:  0.6653,   Adjusted R-squared:  0.6607
```

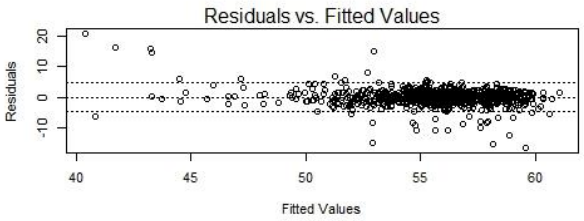
The only statistically significant variables are AGE with a negative effect, AGE:SEXmale with a positive effect when compared to females, SEXmale:DIETunhealthy with a positive effect and DIETunhealthy:SPORT with a positive effect. The adjusted R-Squared is much better here with a value of 0.6607.

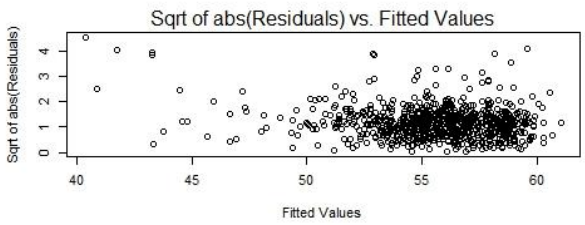
Residuals











Variables 44, 242, 364, 385, 405, 479, 500 and 645 are outliers.