Methods

Zayed Shahjahan

12/9/2020

We will now examine the methods used in the analysis of the Obesity data. First we start by examining the first 5 entries

```
##   LocationAbbr             Income perc_strength propstrength adjstrength
## 1           AK  $15,000 - $24,999          38.6   0.13217139    5.101816
## 2           AK  $25,000 - $34,999          30.7   0.09513435    2.920625
## 3           AK  $35,000 - $49,999          34.6   0.13144517    4.548003
## 4           AK  $50,000 - $74,999          31.5   0.18155410    5.718954
## 5           AK $75,000 or greater          35.4   0.45969499   16.273203
## 6           AL  $15,000 - $24,999          21.0   0.24905057    5.230062
##   perc_cardio propcardio adjcardio perc_obese prop_obese  adjobese perc_excer
## 1        51.7 0.13062731  6.753432       31.0 0.13894812  4.307392       24.8
## 2        55.8 0.09372694  5.229963       26.0 0.09559346  2.485430       25.8
## 3        59.6 0.13173432  7.851365       31.7 0.13077470  4.145558       22.8
## 4        59.1 0.18339483 10.838635       29.0 0.18194741  5.276475       23.9
## 5        63.9 0.46051661 29.427011       30.8 0.45273632 13.944279       26.6
## 6        39.3 0.24731183  9.719355       39.3 0.25217897  9.910633       11.7
##   prop_excer  adjexcer perc_fruit prop_fruit  adjfruit perc_veg   prop_veg
## 1  0.1292975  3.206577       43.5 0.13179690  5.733165     27.5 0.13216502
## 2  0.0941704  2.429596       42.2 0.09686712  4.087793     17.3 0.09492515
## 3  0.1315396  2.999103       45.1 0.13323731  6.009003     15.9 0.13253012
## 4  0.1827354  4.367377       38.0 0.18077062  6.869283     14.6 0.18108799
## 5  0.4622571 12.296039       35.9 0.45732805 16.418077     15.3 0.45929171
## 6  0.2482618  2.904663       50.3 0.24595405 12.371489     34.2 0.24532520
##     adjveg vrich  region
## 1 3.634538     0 Western
## 2 1.642205     0 Western
## 3 2.107229     0 Western
## 4 2.643885     0 Western
## 5 7.027163     1 Western
## 6 8.390122     0   South
```

We first fit a model to the unadjusted variables. We use only the numeric variables

```
##
## Call:
## lm(formula = perc_obese ~ perc_strength + perc_cardio + perc_excer +
##     perc_fruit + perc_veg, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1912  -2.0820   0.0282   2.4060   8.2505
##
## Coefficients:
```

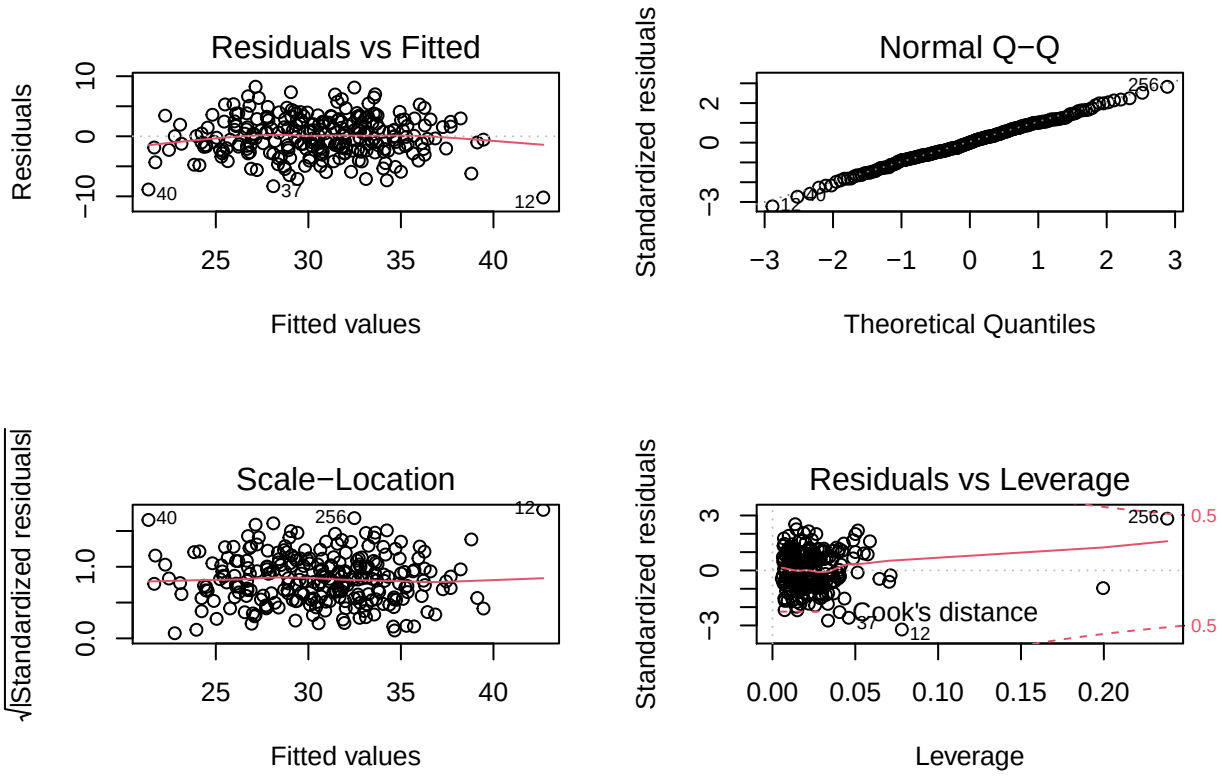
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.89206    4.32651   8.989  < 2e-16 ***
## perc_strength -0.30948    0.09287  -3.333 0.000991 ***
## perc_cardio   -0.15666    0.06103  -2.567 0.010842 *
## perc_excer    -0.08031    0.12686  -0.633 0.527271
## perc_fruit     0.31091    0.04732   6.571 2.89e-10 ***
## perc_veg      -0.09496    0.06004  -1.582 0.114994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.295 on 250 degrees of freedom
## Multiple R-squared:  0.5737, Adjusted R-squared:  0.5652
## F-statistic: 67.29 on 5 and 250 DF,  p-value: < 2.2e-16
```

The VIF is:

```
## perc_strength   perc_cardio    perc_excer    perc_fruit      perc_veg
##      5.652806      4.487290      8.686051      1.758501      3.091466
```

This is to be expected as perc_excer is strongly correlated with perc_strength and perc_cardio.

We will also perform some regression diagnostics on this model



Notice that our added random data point has been flagged as a potential outlier. This model is to serve as our reference model. The population adjusted model with Income and State-level dummies is our primary focus.

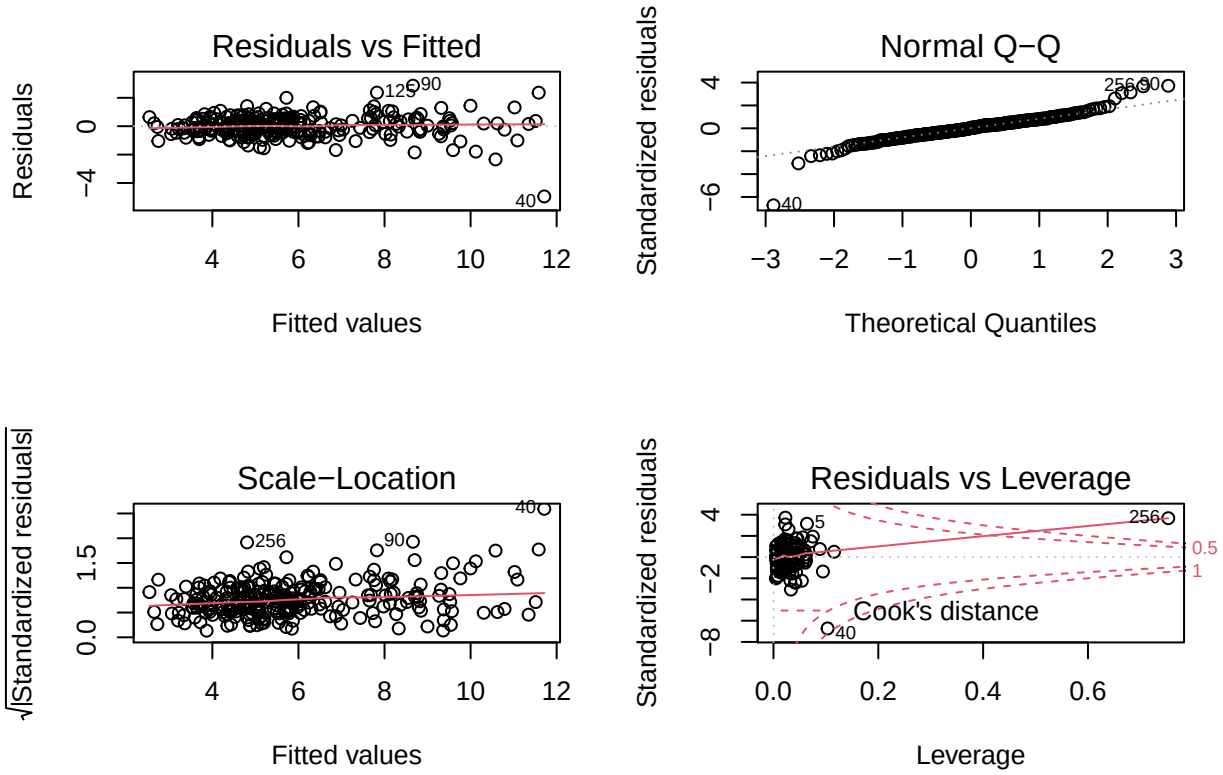
```
##
## Call:
```

```
## lm(formula = adjobese ~ adjstrength + adjcardio + adjexcer +
##     adjfruit + adjveg, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9479 -0.4030  0.0298  0.4414  2.8555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.42629    0.15348   2.777 0.005894 **
## adjstrength -0.09257    0.07009  -1.321 0.187850
## adjcardio    0.16747    0.05014   3.340 0.000965 ***
## adjexcer    -0.18578    0.09194  -2.021 0.044383 *
## adjfruit     0.47934    0.05198   9.221  < 2e-16 ***
## adjveg       0.30581    0.06655   4.595 6.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7765 on 250 degrees of freedom
## Multiple R-squared:  0.8695, Adjusted R-squared:  0.8669
## F-statistic: 333.1 on 5 and 250 DF,  p-value: < 2.2e-16
```

The VIF is:

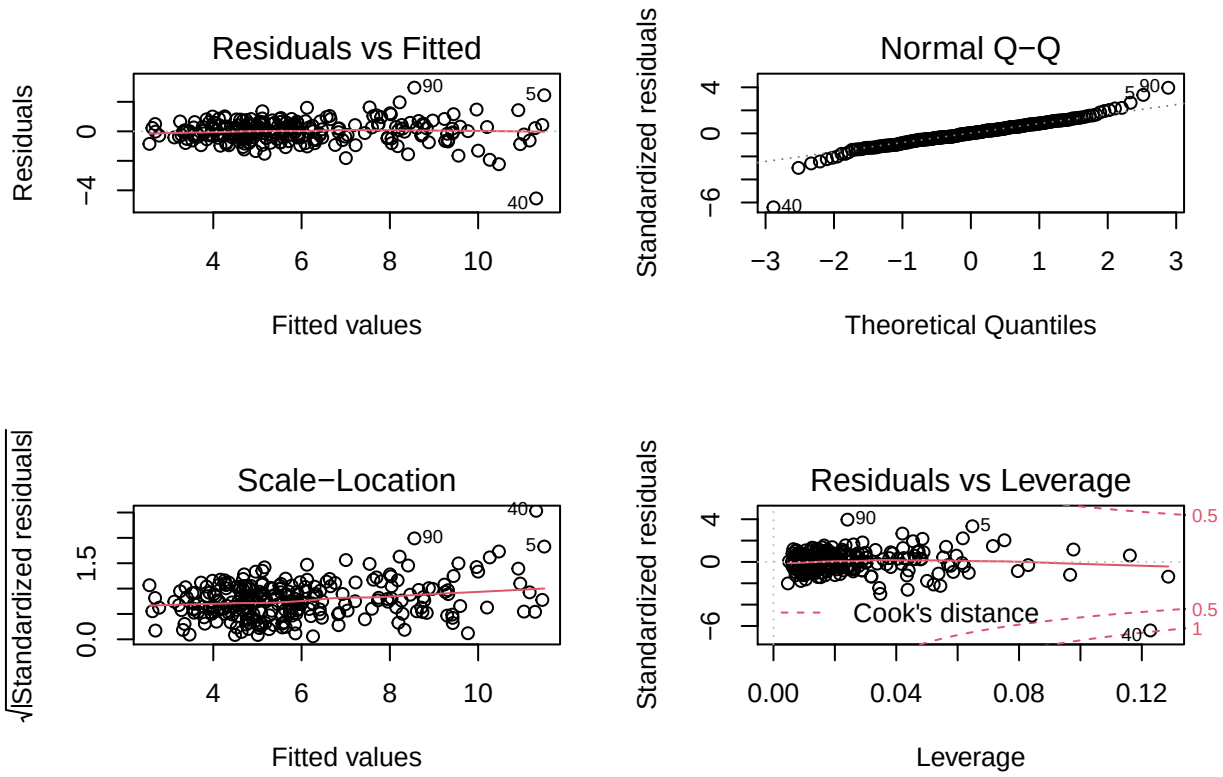
```
## adjstrength   adjcardio    adjexcer    adjfruit      adjveg
##   30.857061   39.151581   31.340168   11.064115    3.981397
```

This is immediately an issue. Multicollinearity is very pronounced in the second model and needs to be addressed.



Again, our added datapoint is flagged as an outlier and it is likely that this is due to picking points at random. It is complicating things so I am going to work on dataset without this point

```
##
## Call:
## lm(formula = adjobese ~ adjstrength + adjcardio + adjexcer +
##     adjfruit + adjveg, data = df[-256, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5498 -0.3979  0.0102  0.4298  2.9600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.32932    0.15178   2.170   0.0310 *
## adjstrength  0.24486    0.11265   2.174   0.0307 *
## adjcardio    0.22860    0.05149   4.440 1.35e-05 ***
## adjexcer    -0.71465    0.16654  -4.291 2.55e-05 ***
## adjfruit     0.44240    0.05160   8.573 1.08e-15 ***
## adjveg       0.28186    0.06517   4.325 2.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7567 on 249 degrees of freedom
## Multiple R-squared:  0.8765, Adjusted R-squared:  0.874
## F-statistic: 353.5 on 5 and 249 DF,  p-value: < 2.2e-16
```



Without point 256, the only other potential outlier is point 40, this corresponds to the $75,000 or greater income bracket of the District of Columbia.

In order to deal with the multicollinearity without droping variables, I have decided to combine the three physical activity related variables into one

Below is the summary for this new variable

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.810   5.125   6.383   7.595   7.783  23.109
```

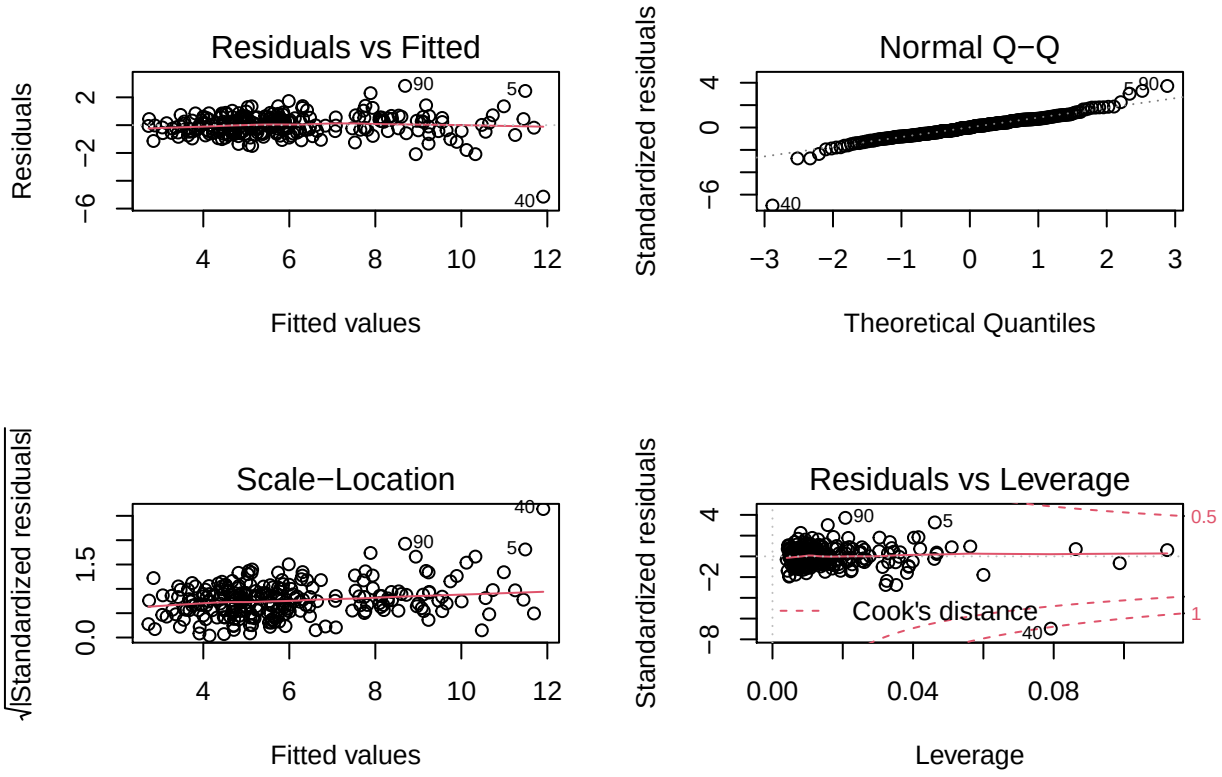
Refitting the model with this new variable yields the following:

```
##
## Call:
## lm(formula = adjobese ~ phys + adjfruit + adjveg, data = df[-256,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1412 -0.4299  0.0038  0.4553  2.8228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.55441    0.14684   3.776 0.000199 ***
## phys         0.12366    0.03298   3.750 0.000220 ***
## adjfruit     0.33791    0.05503   6.140 3.20e-09 ***
## adjveg       0.41641    0.06330   6.578 2.75e-10 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7686 on 251 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8701
## F-statistic:   568 on 3 and 251 DF,  p-value: < 2.2e-16
```

The VIF is:

```
##      phys  adjfruit    adjveg
##  6.981947 12.649965  3.676246
```



When taking out the outlier we observe the following:

```
##
## Call:
## lm(formula = adjobese ~ phys + adjfruit + adjveg, data = df[-c(256,
##     40), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34966 -0.44730  0.00928  0.44341  2.61050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.46866    0.13259   3.535 0.000486 ***
## phys         0.14299    0.02978   4.802 2.71e-06 ***
## adjfruit     0.35632    0.04957   7.187 7.60e-12 ***
## adjveg       0.37124    0.05726   6.484 4.74e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6915 on 250 degrees of freedom
## Multiple R-squared:  0.8964, Adjusted R-squared:  0.8952
## F-statistic:   721 on 3 and 250 DF,  p-value: < 2.2e-16
```

We will now perform a robust regression with Huber's psi function to check whether the coefficients differ significantly

```
##
## Call: rlm(formula = adjobese ~ phys + adjfruit + adjveg, data = df[-c(256,
##     40), ])
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.365202 -0.452185  0.009728  0.445682  2.596608
##
## Coefficients:
##             Value  Std. Error t value
## (Intercept) 0.4591 0.1218     3.7683
## phys        0.1436 0.0274     5.2472
## adjfruit    0.3586 0.0456     7.8713
## adjveg      0.3683 0.0526     6.9992
##
## Residual standard error: 0.6647 on 250 degrees of freedom
```

From this we conclude that our non-robust model is adequate. And we will build on it.

To address the multicollinearity concern for the lack of fruit and vegetable consumption we will perform the same procedure on these to variables

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.359   4.597   5.568   6.059   7.063  13.067
```

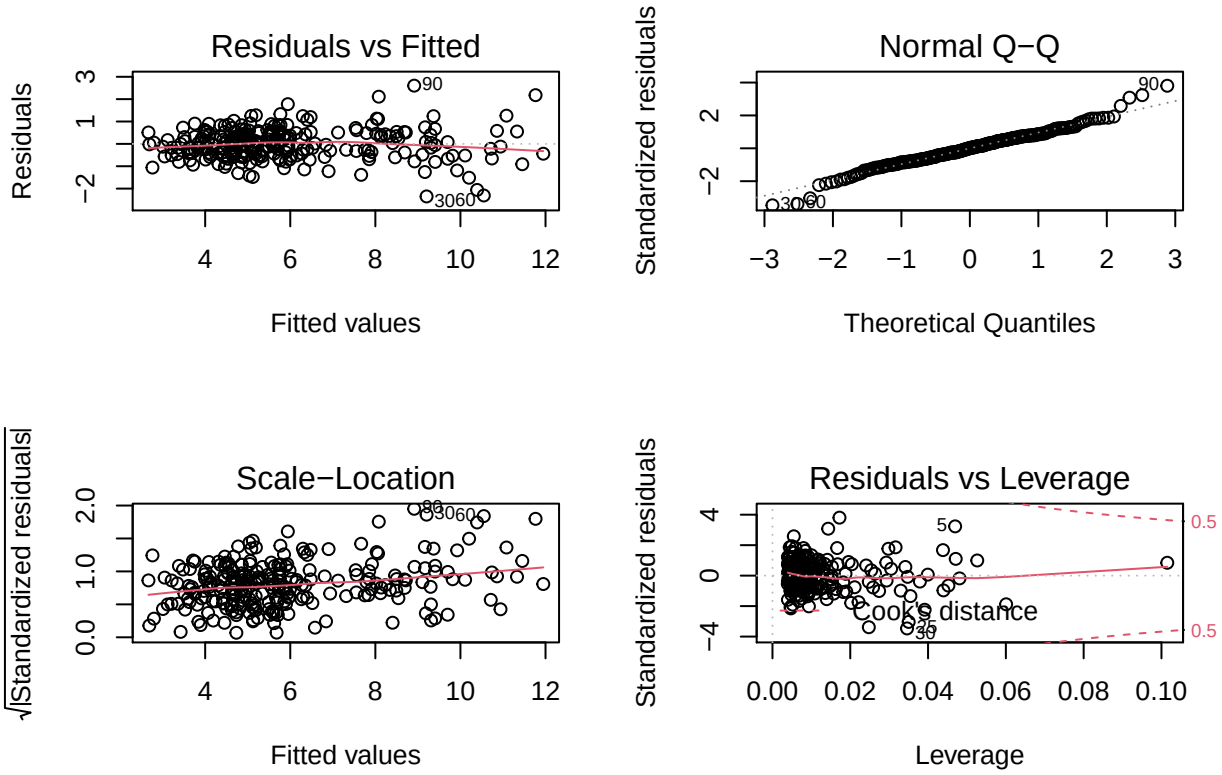
Using this adjustment yields the following model

```
##
## Call:
## lm(formula = adjobese ~ phys + nutri, data = df[-c(256, 40),
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34997 -0.44386  0.00904  0.44280  2.60318
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.47126    0.13116   3.593 0.000393 ***
## phys         0.13982    0.02068   6.760 9.62e-11 ***
## nutri        0.72634    0.03593  20.217  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6902 on 251 degrees of freedom
## Multiple R-squared:  0.8964, Adjusted R-squared:  0.8956
## F-statistic:  1086 on 2 and 251 DF,  p-value: < 2.2e-16
```

```
##     phys    nutri
```

```
## 3.189118 3.189118
```

We can now be assured that our estimates can be trusted

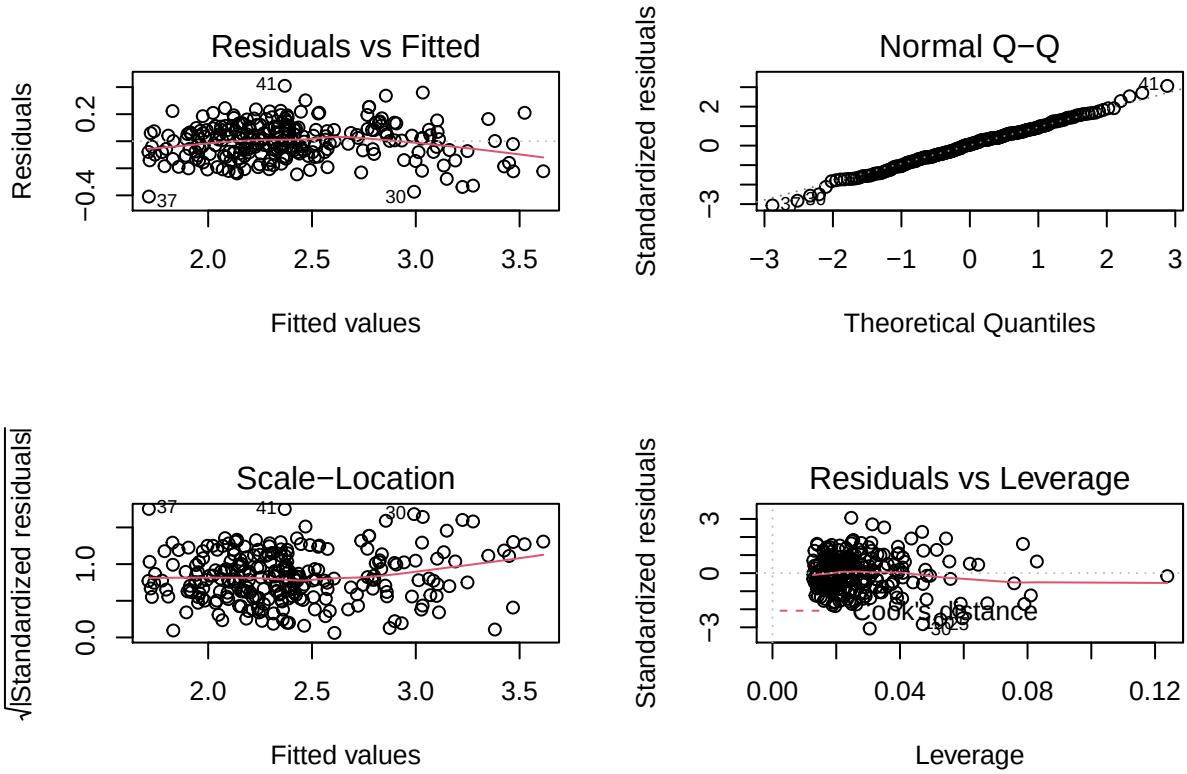


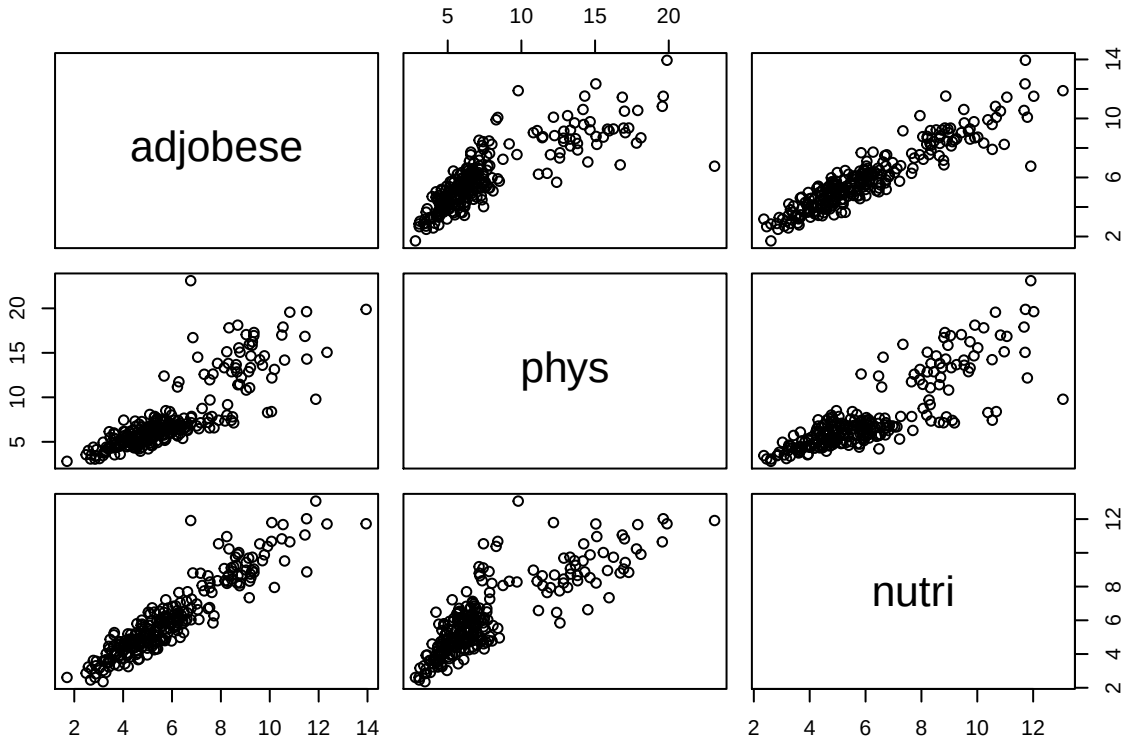
If we add the vrich dummy variable to the model we obtain:

```
##
## Call:
## lm(formula = adjobese ~ phys + nutri + vrich, data = df[-c(256,
##     40), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40794 -0.41897  0.01276  0.40934  2.67995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.22611    0.17591   1.285    0.200
## phys         0.19543    0.03378   5.786 2.15e-08 ***
## nutri        0.71455    0.03614  19.770  < 2e-16 ***
## vrich       -0.52214    0.25167  -2.075    0.039 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6857 on 250 degrees of freedom
## Multiple R-squared:  0.8981, Adjusted R-squared:  0.8969
## F-statistic: 734.8 on 3 and 250 DF,  p-value: < 2.2e-16
```

When controlling for region and adjusting for the u-shaped pattern of the residuals:

```
##
## Call:
## lm(formula = sqrt(adjobese) ~ phys + nutri + vrich + region,
##     data = df[-c(256, 40), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40919 -0.08231  0.00711  0.08606  0.40955
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.285745   0.039243  32.764  < 2e-16 ***
## phys              0.060109   0.007344   8.185 1.45e-14 ***
## nutri             0.117194   0.008526  13.745  < 2e-16 ***
## vrich            -0.213661   0.050507  -4.230 3.29e-05 ***
## regionNortheast -0.046626   0.025437  -1.833  0.06800 .
## regionSouth       0.062854   0.022773   2.760  0.00621 **
## regionWestern    -0.116762   0.027866  -4.190 3.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1354 on 247 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8987
## F-statistic: 375.1 on 6 and 247 DF,  p-value: < 2.2e-16
```





The physical activity cluster is more pronounced after the variables are merged into one. The lack nutrition has the greatest linear effect on Obesity rates and this demands some investigation

```
##
## Call:
## lm(formula = sqrt(adjobese) ~ nutri + vrich + region, data = df[-c(256,
##     40), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48886 -0.10471  0.00139  0.09771  0.40811
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.371281   0.042562  32.218  < 2e-16 ***
## nutri            0.165966   0.006862  24.186  < 2e-16 ***
## vrich            0.108912   0.035544   3.064  0.00242 **
## regionNortheast -0.011026   0.028200  -0.391  0.69613
## regionSouth      0.033132   0.025297   1.310  0.19151
## regionWestern   -0.053652   0.030131  -1.781  0.07620 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1523 on 248 degrees of freedom
## Multiple R-squared:  0.8743, Adjusted R-squared:  0.8717
## F-statistic: 344.9 on 5 and 248 DF,  p-value: < 2.2e-16
```

The change in signs on the vrich variables gives credence to the idea that physical activity and Income move in the same direction and this latent influence is reflected in our model that includes both of these variables.

```
##
## Call: rlm(formula = sqrt(adjobese) ~ phys + nutri + vrich + region,
##     data = df[-c(256, 40), ])
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.403008 -0.084481  0.004666  0.085248  0.401241
##
## Coefficients:
##                 Value   Std. Error t value
## (Intercept)      1.2643  0.0408    30.9745
## phys             0.0616  0.0076     8.0631
## nutri            0.1196  0.0089    13.4910
## vrich           -0.2305  0.0525    -4.3879
## regionNortheast -0.0419  0.0265    -1.5826
## regionSouth      0.0557  0.0237     2.3502
## regionWestern   -0.1053  0.0290    -3.6334
##
## Residual standard error: 0.1266 on 247 degrees of freedom
```

Therefore our final model is one where the square root of adjusted obesity percentage is the target variable and the regressors are: phys (physical activity), nutri(lack of nutrition), vrich(dummy variable: 1 if income 75,000 or greater, 0 otherwise), region(factor)