Predictive Modeling of Covid Recovery Rates with Protein Sources

Meher Ivatury

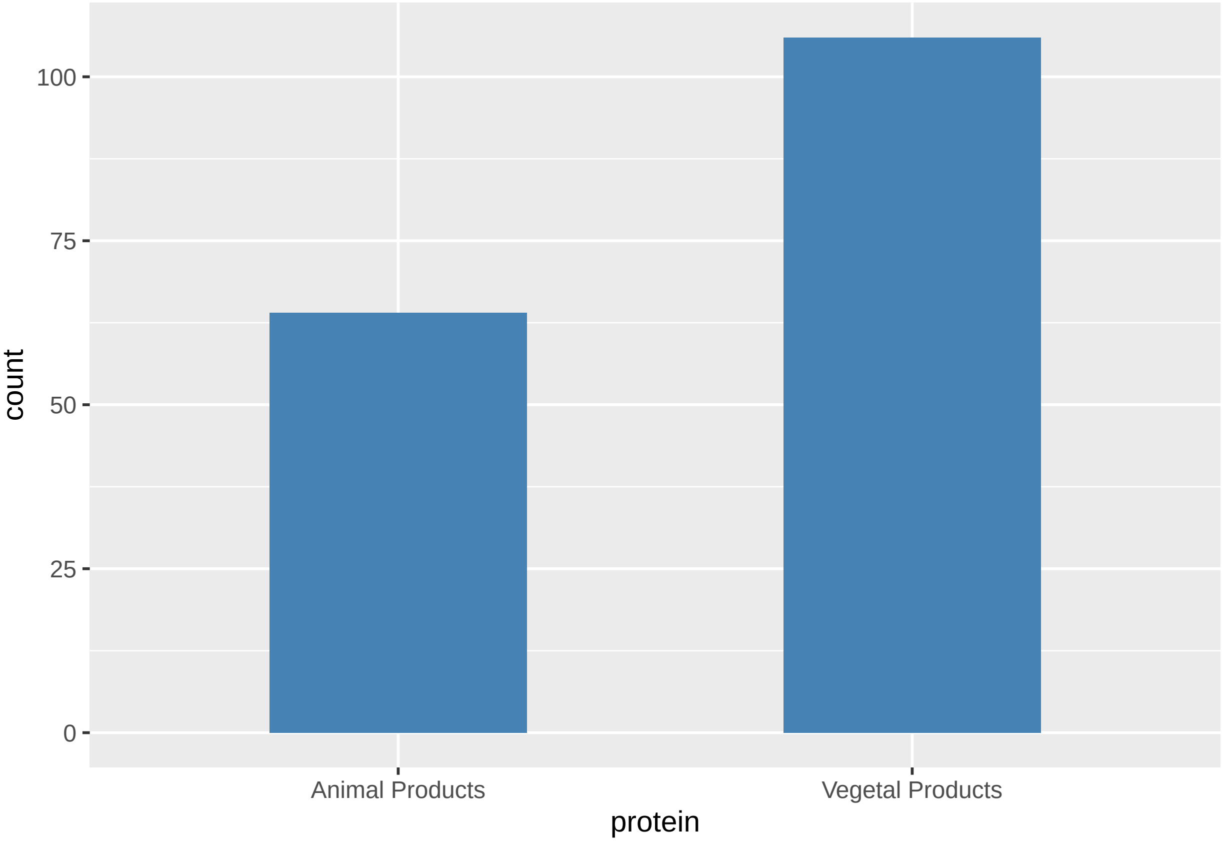
- Title Page
- Introduction and Data Description
- Exploratory Data Analysis
- Methodology
- Discussion and Conclusion
- Presentation and Organization

The goal of this project is to determine whether or not there is a relationship between the how a population intakes their protein and what percent of the population recovers from Covid-19. The data set contains data on 170 countries and tracks 32 variables. The majority of these variables are forms of protein intake such as animal products, eggs, starchy roots, etc and another important variable is recovered cases which is a percentage describing what percent of confirmed cases had recovered from the Covid-19 virus. The final step will be to create a regresion model that will attempt to predict the percent of recovered cases from protein intake figures.

Note that French Polynesia, Kiribati, North Korea, Myanmar, New Caledonia, and Turkmenistan do not have available data for recovered cases. Belgium, Serbia, Sweden, and The United States of America all have a 0 percent recovery rate as well.

```
## [1] 64
```

```
## [1] 106
```



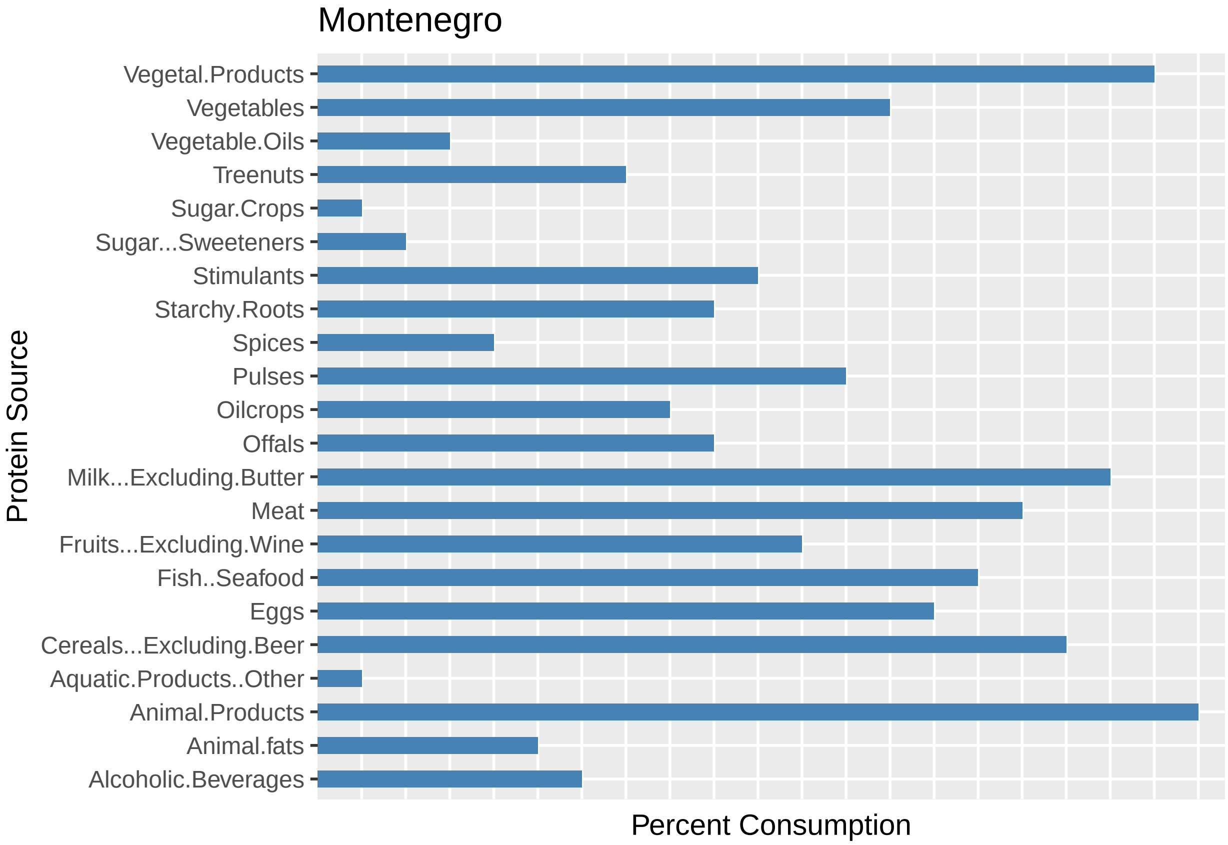
The majority of the 170 countries get most of their protein from vegetal products. 106 countries are from vegetal products and 64 are from animal products. After taking a look at the data it is clear that countries from Asia have a more vegetal based diet and European countries will have a more animal product based diet.

```
## [1] Montenegro Czechia    Luxembourg Slovenia   Georgia    Panama
## 170 Levels: Afghanistan Albania Algeria Angola Antigua and Barbuda ... Zimbabwe
```

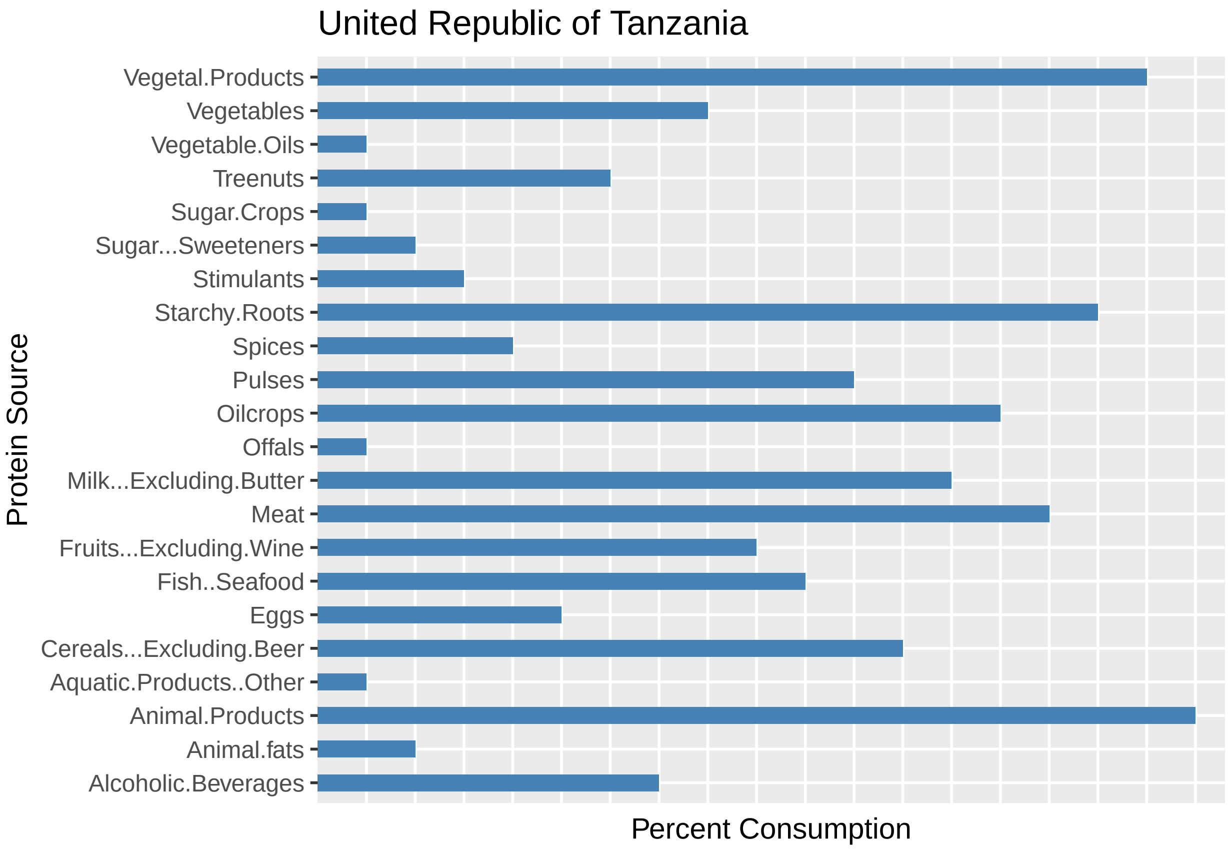
```
## [1] Vietnam                          Solomon Islands
## [3] Samoa                            Lao People's Democratic Republic
## [5] Vanuatu                          United Republic of Tanzania
## 170 Levels: Afghanistan Albania Algeria Angola Antigua and Barbuda ... Zimbabwe
```

The top five countries with the highest recovery rates are Montenegro, Czechia, Luxembourg, Slovenia, Georgia, Panama

The bottom five countries with the lowest recovery rates are Vietnam, Solomon Islands, Samoa, Lao People's Democratic Republic Vanuatu, United Republic of Tanzania



This is a bar plot of the protein intakes of Montenegro, the country with the highest recovery rate. As you can see the majority of their protein comes from animal products, vegetal products, and milk products.



This is a bar plot of the protein intakes of the population of the United Republic of Tanzania. The majority of their protein comes from vegetal products and cereals.

```
r_protein <- subset(protein, select = -c(Unit..all.except.Population., Confirmed, Deaths, Active, Countr
```

Initial Data Manipulation

- Remove columns "Confirmed", "Deaths", and "Active" because of multicollinearity.
- Removed "Unit..all.except.Population." because there is no information in this column.
- Removed column "Country" because this model should be able to predict regardless of which country is inputted.
- Change the factor level "<2.5" to "2.5" in order to turn Undernourished to a numeric variable.
- Removed NA's in the data set

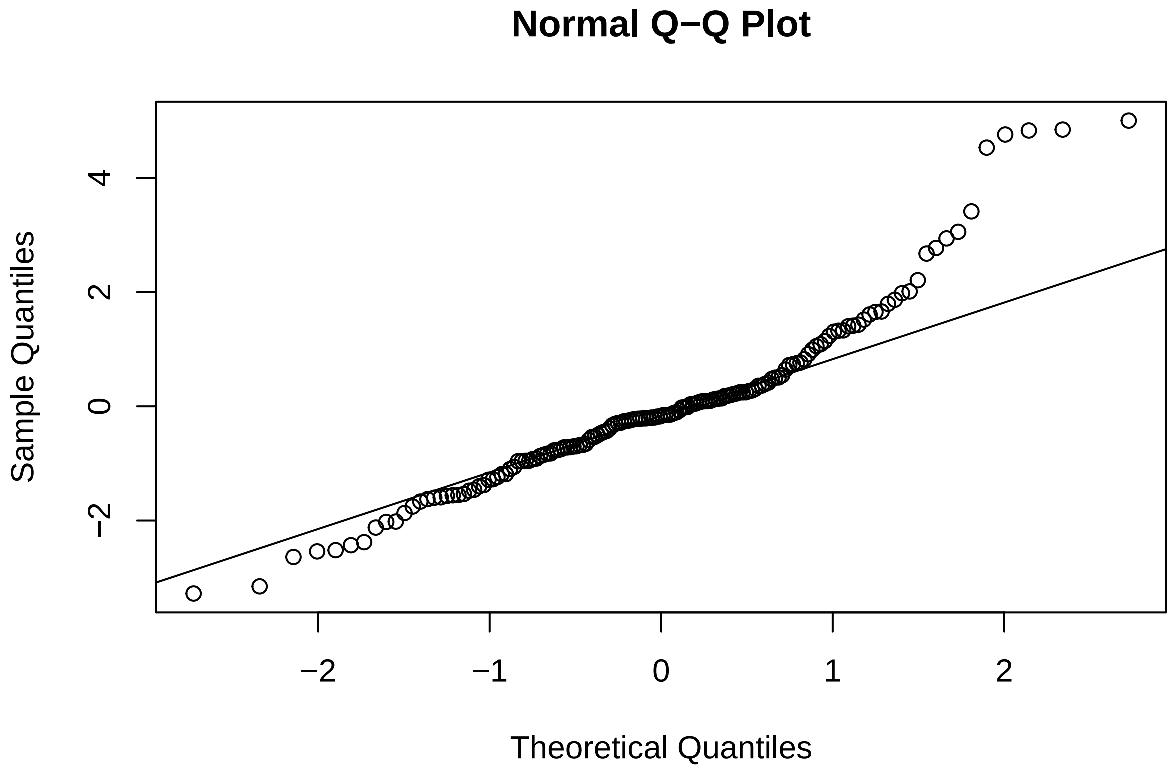
OLS Regression

Fit a multiple linear regression model

```
##
## Call:
## lm(formula = Recovered ~ ., data = r_protein)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2784 -0.8340 -0.1672  0.5036  5.0051
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.457e+05  1.049e+05   2.342   0.0207 *
## Alcoholic.Beverages       -2.438e+03  1.048e+03  -2.327   0.0215 *
## Animal.Products           -2.473e+03  1.051e+03  -2.352   0.0202 *
## Animal.fats               -2.439e+03  1.048e+03  -2.328   0.0215 *
## Aquatic.Products..Other   -2.445e+03  1.048e+03  -2.333   0.0212 *
## Cereals...Excluding.Beer  -2.439e+03  1.048e+03  -2.328   0.0215 *
## Eggs                      -2.441e+03  1.048e+03  -2.330   0.0214 *
## Fish..Seafood             -2.442e+03  1.048e+03  -2.330   0.0213 *
## Fruits...Excluding.Wine   -2.439e+03  1.048e+03  -2.328   0.0215 *
## Meat                      -2.442e+03  1.048e+03  -2.330   0.0213 *
## Milk...Excluding.Butter   -2.441e+03  1.048e+03  -2.330   0.0213 *
## Offals                    -2.442e+03  1.048e+03  -2.330   0.0213 *
## Oilcrops                  -2.439e+03  1.048e+03  -2.328   0.0214 *
## Pulses                    -2.439e+03  1.048e+03  -2.328   0.0214 *
## Spices                    -2.439e+03  1.048e+03  -2.328   0.0215 *
## Starchy.Roots             -2.439e+03  1.048e+03  -2.328   0.0214 *
## Stimulants                -2.439e+03  1.048e+03  -2.328   0.0215 *
## Sugar.Crops               -2.442e+03  1.047e+03  -2.332   0.0213 *
## Sugar...Sweeteners        -2.432e+03  1.048e+03  -2.322   0.0218 *
## Treenuts                  -2.439e+03  1.048e+03  -2.328   0.0215 *
## Vegetal.Products          -2.476e+03  1.051e+03  -2.355   0.0200 *
## Vegetable.Oils            -2.457e+03  1.047e+03  -2.346   0.0205 *
## Vegetables                -2.439e+03  1.048e+03  -2.328   0.0215 *
## Miscellaneous             -2.438e+03  1.047e+03  -2.328   0.0215 *
## Obesity                    3.141e-02  2.295e-02   1.369   0.1734
## Undernourished            -4.290e-03  5.225e-03  -0.821   0.4132
## Population                -3.936e-10  8.900e-10  -0.442   0.6590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.623 on 129 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.4209, Adjusted R-squared:  0.3042
## F-statistic: 3.606 on 26 and 129 DF,  p-value: 6.884e-07
```

Here we see that we have an r-squared of .4209. Variables that are not significant are Obesity, Undernourshied, and Population.

```
qqnorm(model1$residuals)
qqline(model1$residuals)
```

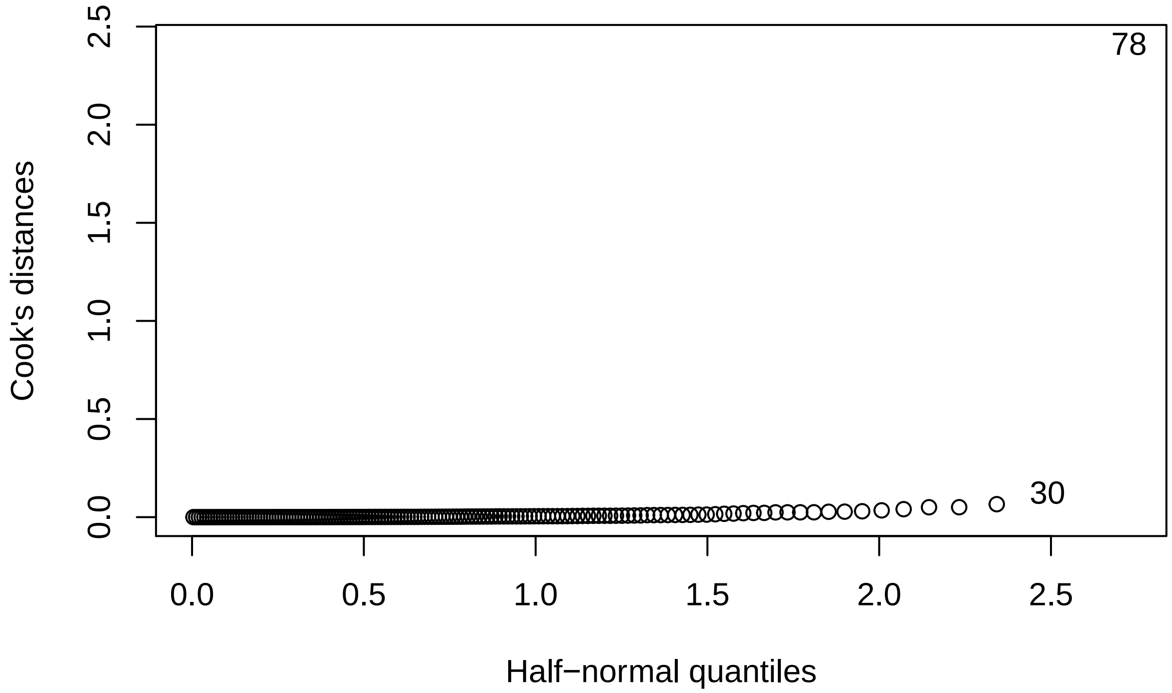


The QQ plot shows outliers that do not follow the normality assumption. The residuals may be correlated as they do not fan out.

```
##      103       74      119       39       56      124      109       92
## 3.357061 3.314111 3.252144 3.243632 3.029870 2.241157 2.207085 2.175243
##      160      139
## 2.041445 1.983999
```

No highly influential points because no value has a leverage greater than 3.712386

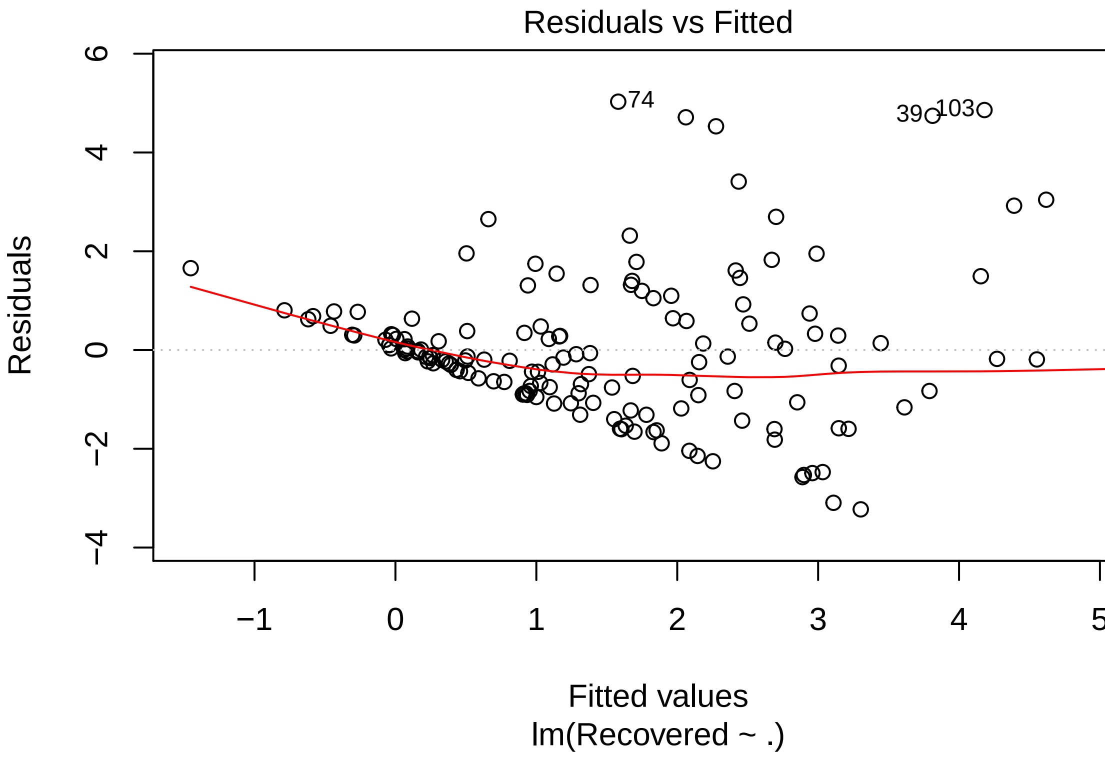
```
## [1] 2.41186
```



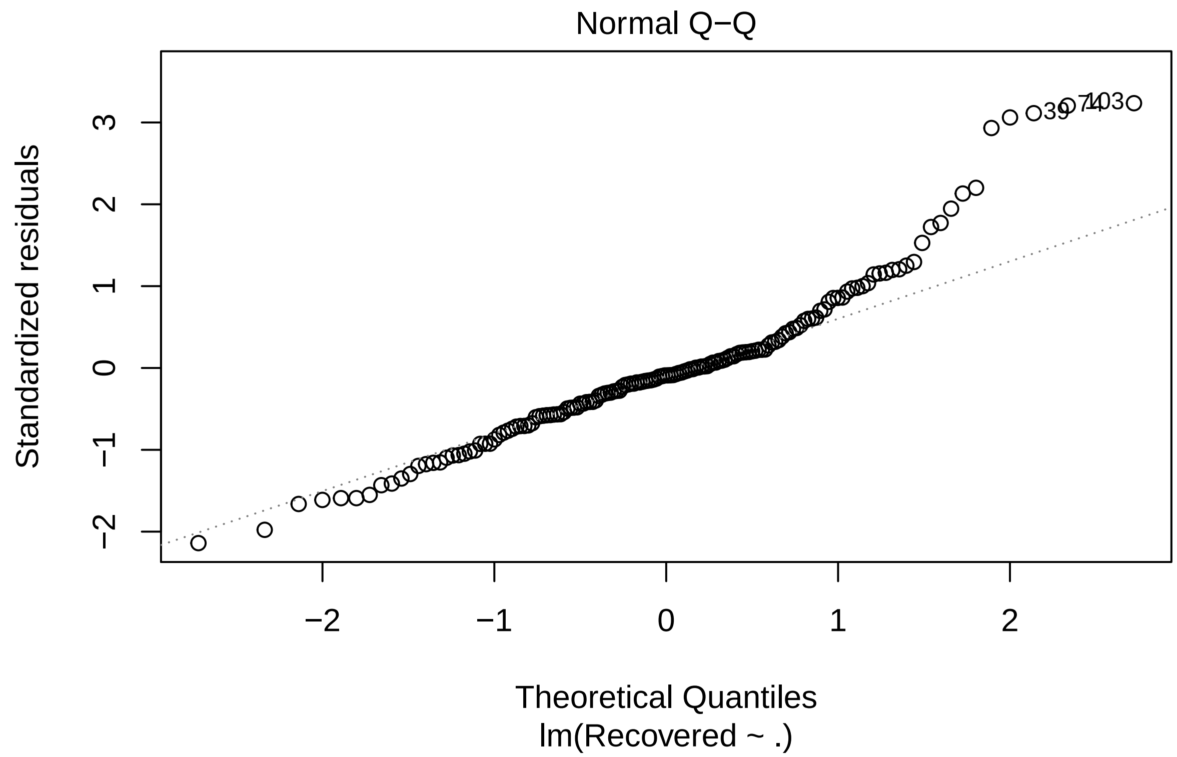
Observations 78 and 30 also need to be removed as they have abnormally large cook's distances

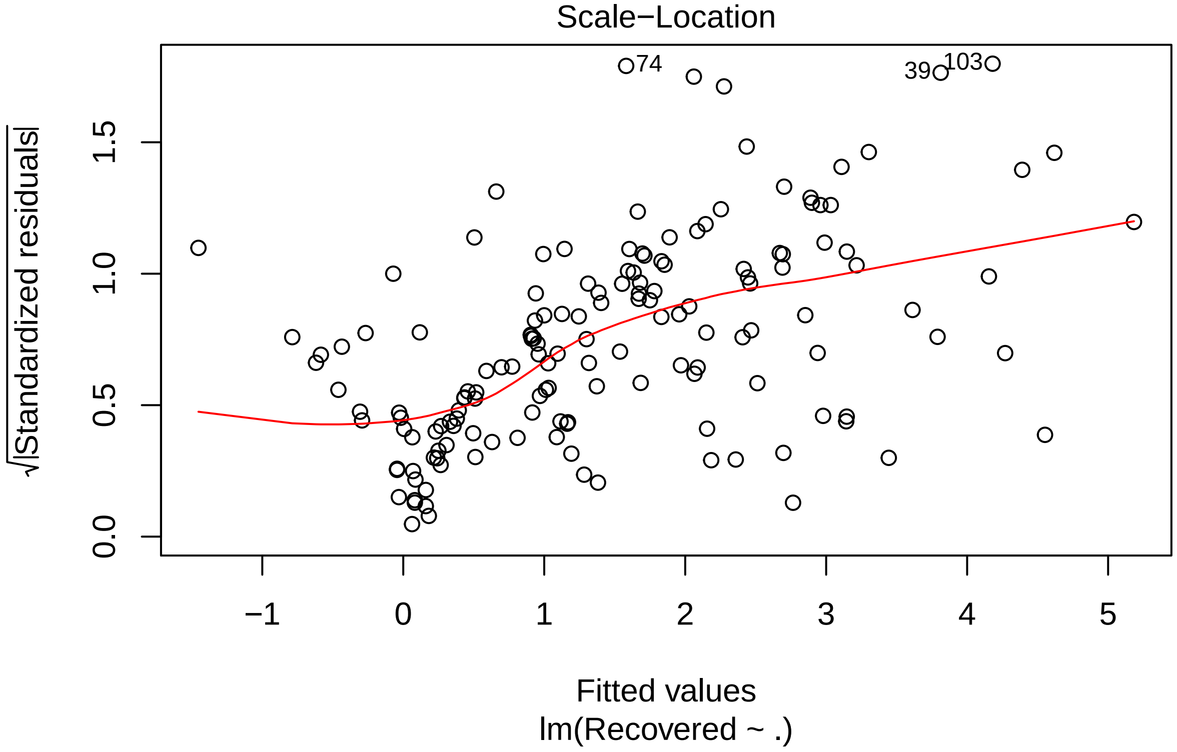
Lets refit the the model

```
##
## Call:
## lm(formula = Recovered ~ ., data = r_protein)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2264 -0.8757 -0.1350  0.5224  5.0283
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.239e+05  1.065e+05   2.102   0.0375 *
## Alcoholic.Beverages      -2.217e+03  1.064e+03  -2.084   0.0392 *
## Animal.Products          -2.262e+03  1.066e+03  -2.122   0.0358 *
## Animal.fats              -2.213e+03  1.065e+03  -2.078   0.0397 *
## Aquatic.Products..Other  -2.218e+03  1.065e+03  -2.083   0.0393 *
## Cereals...Excluding.Beer -2.218e+03  1.064e+03  -2.085   0.0391 *
## Eggs                     -2.215e+03  1.065e+03  -2.080   0.0395 *
## Fish..Seafood            -2.215e+03  1.065e+03  -2.080   0.0395 *
## Fruits...Excluding.Wine  -2.218e+03  1.064e+03  -2.085   0.0391 *
## Meat                     -2.215e+03  1.065e+03  -2.080   0.0395 *
## Milk...Excluding.Butter  -2.215e+03  1.065e+03  -2.080   0.0395 *
## Offals                   -2.215e+03  1.065e+03  -2.080   0.0395 *
## Oilcrops                 -2.218e+03  1.064e+03  -2.085   0.0391 *
## Pulses                   -2.218e+03  1.064e+03  -2.085   0.0391 *
## Spices                   -2.218e+03  1.064e+03  -2.085   0.0391 *
## Starchy.Roots            -2.218e+03  1.064e+03  -2.085   0.0391 *
## Stimulants               -2.218e+03  1.064e+03  -2.084   0.0391 *
## Sugar.Crops              -2.222e+03  1.064e+03  -2.089   0.0387 *
## Sugar...Sweeteners       -2.211e+03  1.064e+03  -2.078   0.0397 *
## Treenuts                 -2.218e+03  1.064e+03  -2.085   0.0391 *
## Vegetal.Products         -2.259e+03  1.067e+03  -2.118   0.0361 *
## Vegetable.Oils           -2.235e+03  1.064e+03  -2.102   0.0376 *
## Vegetables               -2.218e+03  1.064e+03  -2.085   0.0391 *
## Miscellaneous            -2.217e+03  1.064e+03  -2.084   0.0391 *
## Obesity                   3.128e-02  2.303e-02   1.359   0.1767
## Undernourished           -4.748e-03  5.254e-03  -0.904   0.3679
## Population               -4.497e-10  8.927e-10  -0.504   0.6153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.626 on 127 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.4234, Adjusted R-squared:  0.3054
## F-statistic: 3.587 on 26 and 127 DF,  p-value: 8.19e-07
```



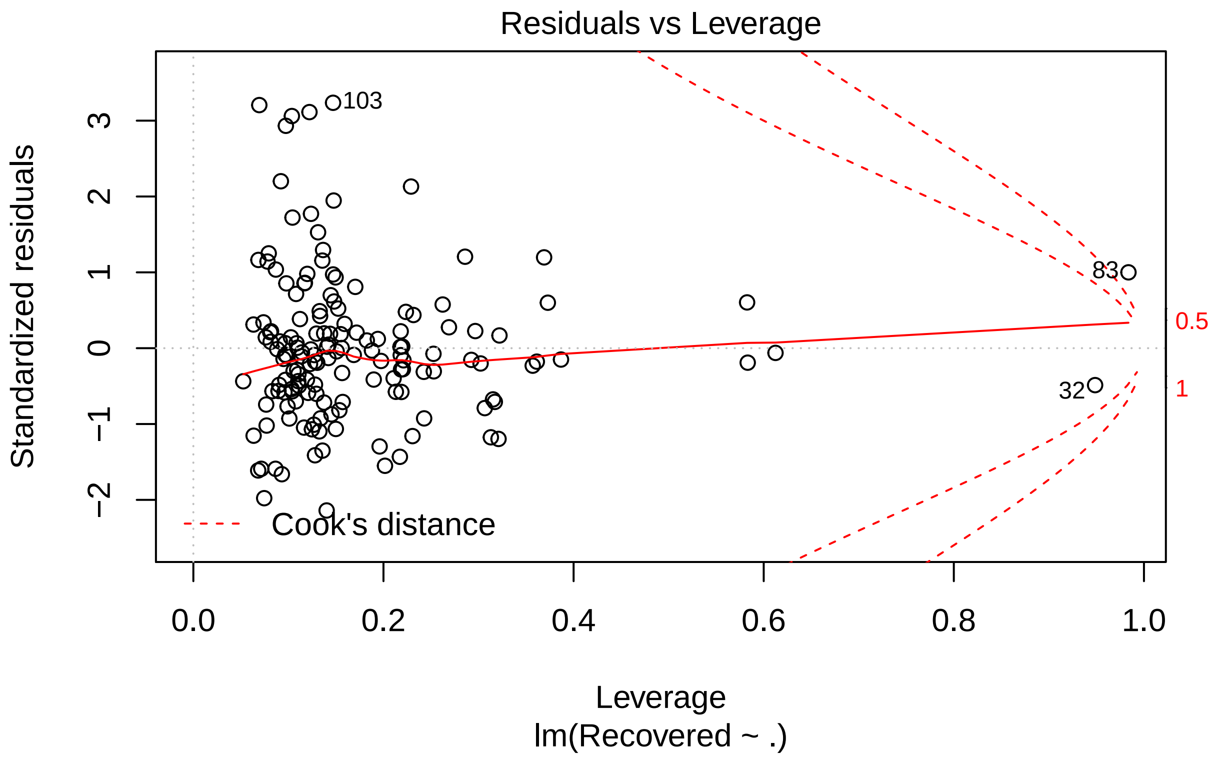
The R-Squared has increased.





```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

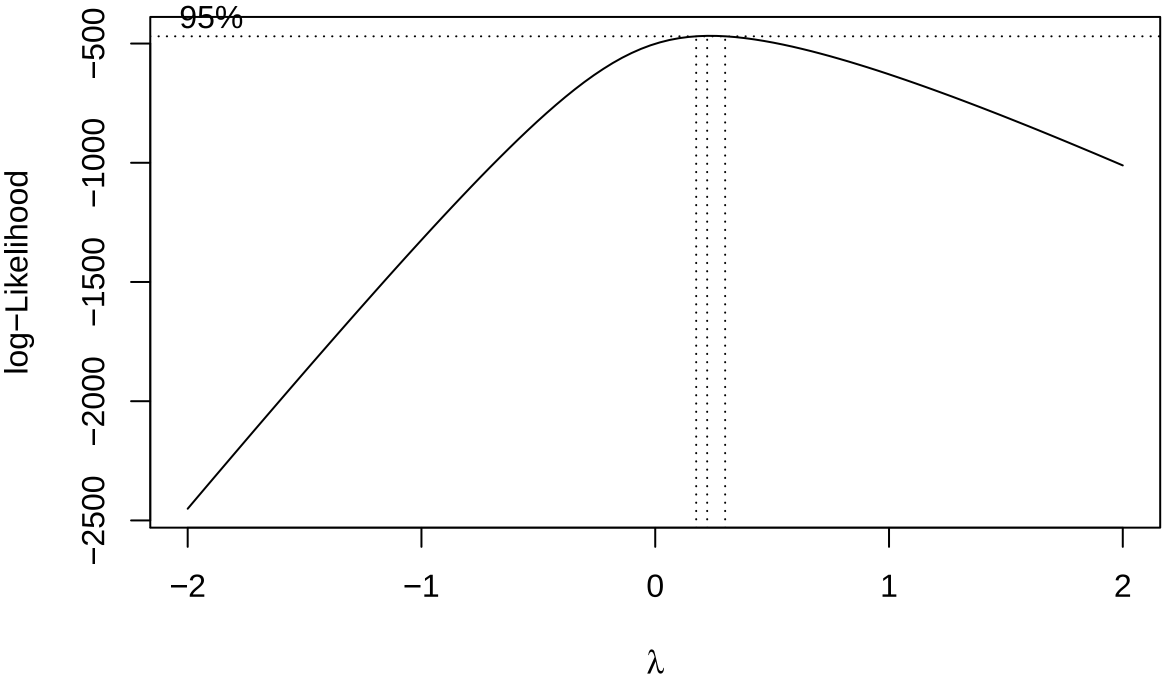


After removing the outliers we can see the normal qqplot does not follow the normal line. Lets try a box-cox tranformation.

```
## Warning: package 'MASS' was built under R version 3.6.2
```

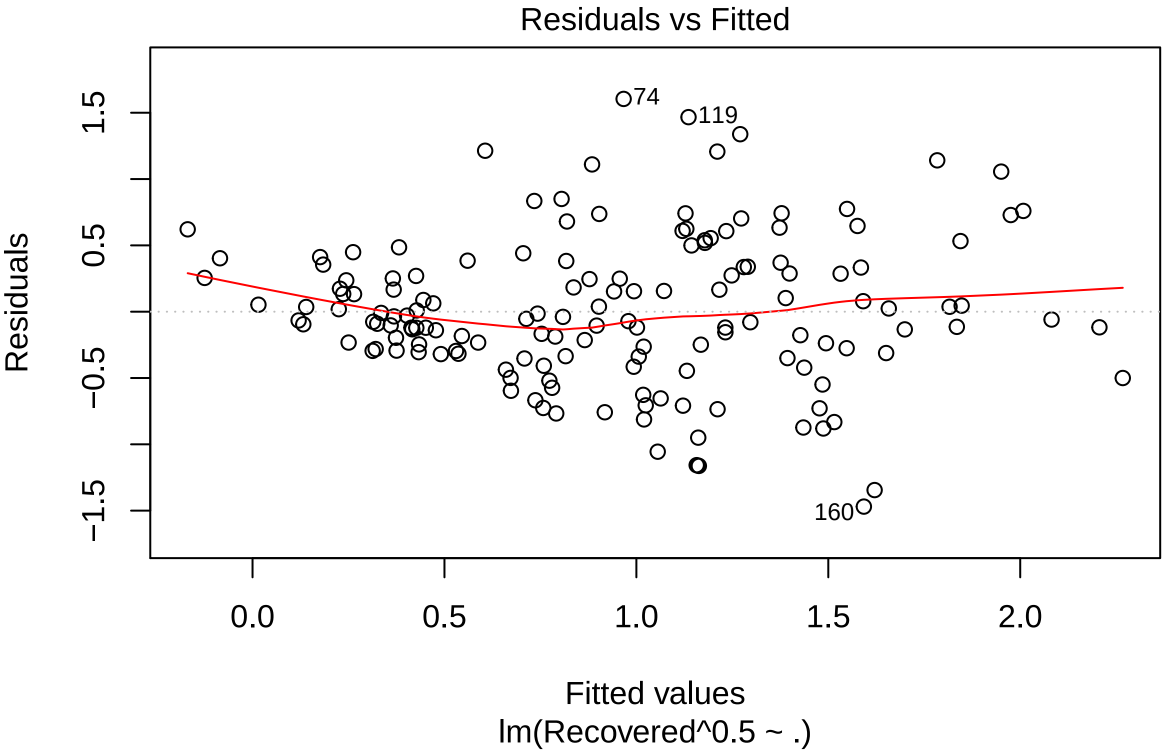
```
##
## Attaching package: 'MASS'
```

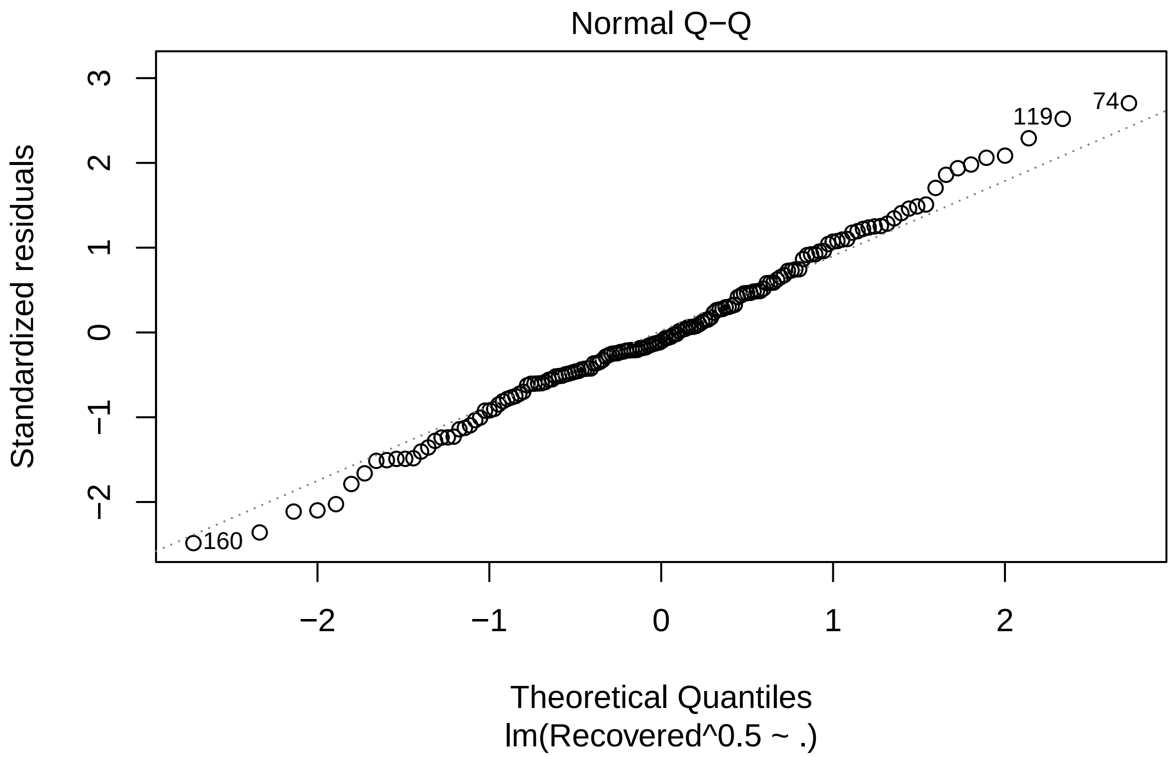
```
## The following object is masked from 'package:dplyr':
##
##     select
```

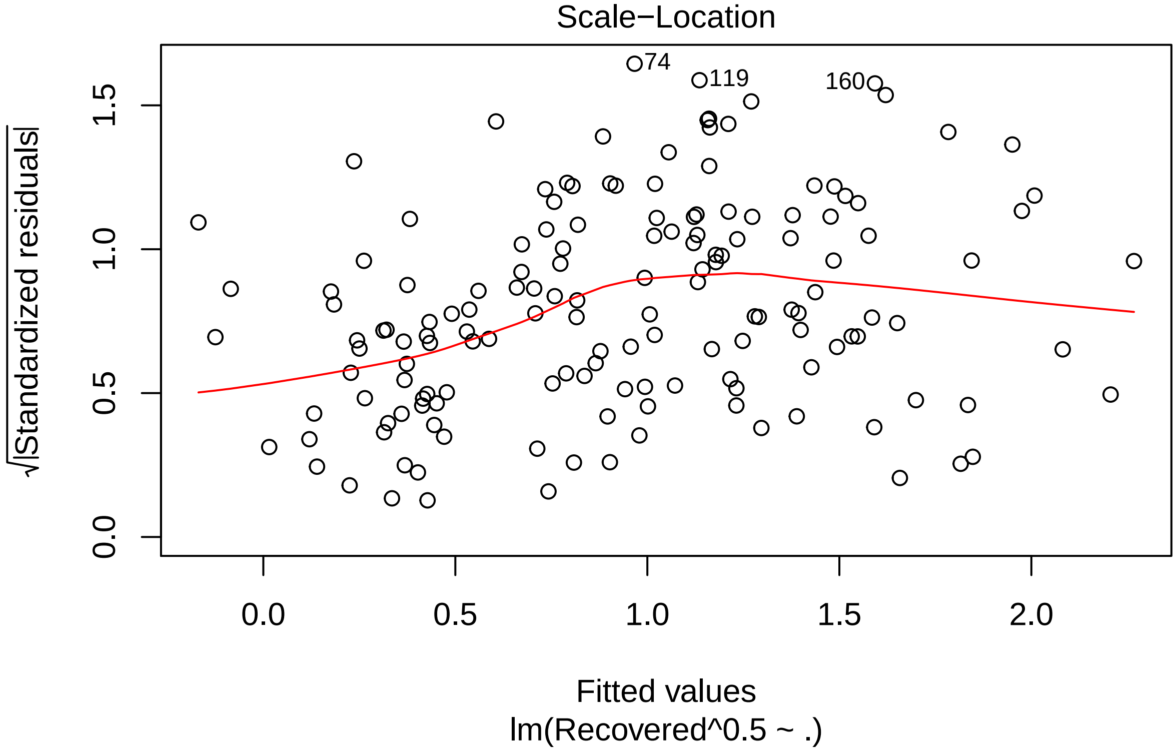


Lambda = .5

```
##
## Call:
## lm(formula = Recovered^0.5 ~ ., data = r_protein)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46954 -0.31009 -0.05645  0.33810  1.60414
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.889e+04  4.027e+04   1.711   0.0896 .
## Alcoholic.Beverages      -6.851e+02  4.022e+02  -1.703   0.0910 .
## Animal.Products          -6.981e+02  4.032e+02  -1.731   0.0858 .
## Animal.fats              -6.789e+02  4.026e+02  -1.686   0.0942 .
## Aquatic.Products..Other  -6.808e+02  4.027e+02  -1.691   0.0933 .
## Cereals...Excluding.Beer -6.856e+02  4.023e+02  -1.704   0.0908 .
## Eggs                     -6.796e+02  4.026e+02  -1.688   0.0939 .
## Fish..Seafood            -6.797e+02  4.026e+02  -1.688   0.0938 .
## Fruits...Excluding.Wine  -6.855e+02  4.023e+02  -1.704   0.0908 .
## Meat                     -6.797e+02  4.026e+02  -1.688   0.0938 .
## Milk...Excluding.Butter  -6.797e+02  4.026e+02  -1.688   0.0938 .
## Offals                   -6.797e+02  4.026e+02  -1.688   0.0938 .
## Oilcrops                 -6.856e+02  4.023e+02  -1.704   0.0907 .
## Pulses                   -6.856e+02  4.023e+02  -1.704   0.0908 .
## Spices                   -6.855e+02  4.023e+02  -1.704   0.0908 .
## Starchy.Roots            -6.856e+02  4.023e+02  -1.704   0.0908 .
## Stimulants               -6.855e+02  4.023e+02  -1.704   0.0908 .
## Sugar.Crops              -6.875e+02  4.022e+02  -1.709   0.0898 .
## Sugar...Sweeteners       -6.832e+02  4.022e+02  -1.698   0.0919 .
## Treenuts                 -6.855e+02  4.023e+02  -1.704   0.0908 .
## Vegetal.Products         -6.922e+02  4.034e+02  -1.716   0.0886 .
## Vegetable.Oils           -6.926e+02  4.022e+02  -1.722   0.0875 .
## Vegetables               -6.855e+02  4.022e+02  -1.704   0.0908 .
## Miscellaneous            -6.852e+02  4.022e+02  -1.704   0.0909 .
## Obesity                   1.482e-02  8.707e-03   1.702   0.0911 .
## Undernourished           -1.758e-03  1.987e-03  -0.885   0.3778
## Population               -2.969e-10  3.375e-10  -0.880   0.3807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6148 on 127 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.4661, Adjusted R-squared:  0.3568
## F-statistic: 4.265 on 26 and 127 DF,  p-value: 1.86e-08
```

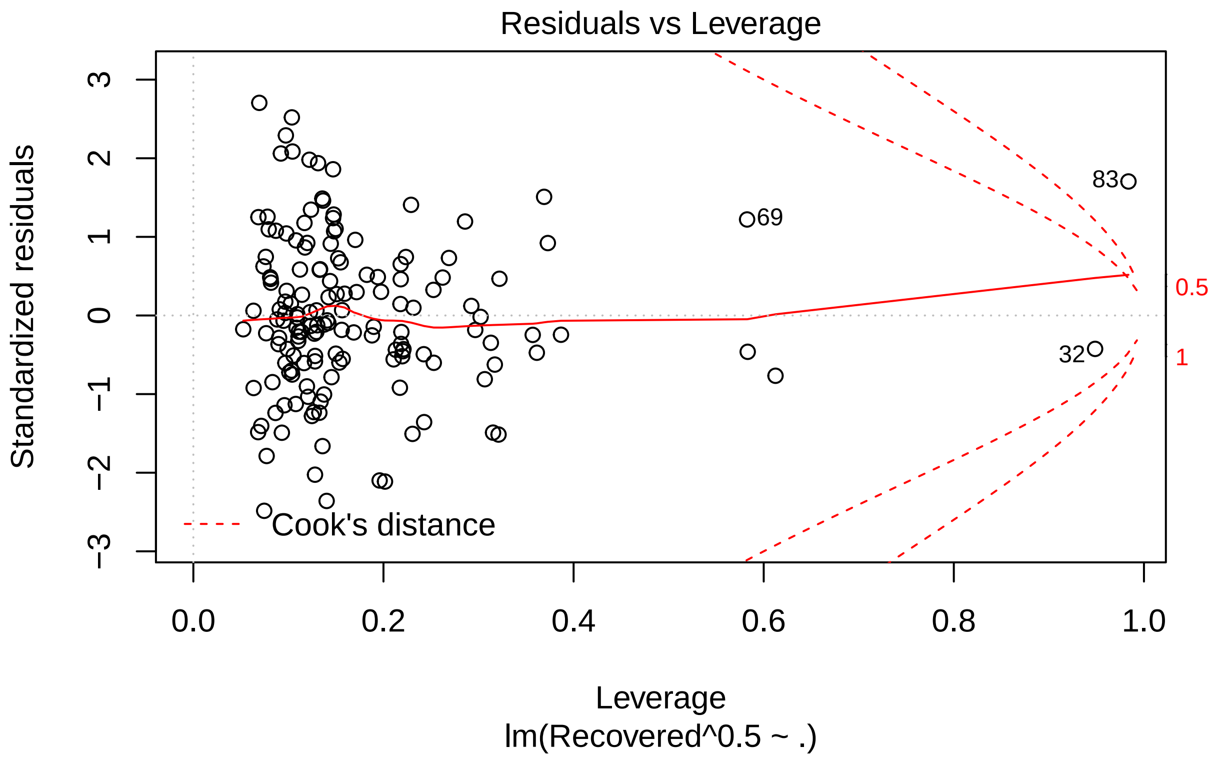






```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



After applying the transformation the qqplot fits much better as well as the residuals seem to spread out more.

After removing the outliers and applying transformations the r-squared improved from 0.4209 to 0.4661 Undernourished and Population have very large P-values therefore it may be better to remove them from the model.

```
## Analysis of Variance Table
##
## Model 1: Recovered^0.5 ~ (Alcoholic.Beverages + Animal.Products + Animal.fats +
##     Aquatic.Products..Other + Cereals...Excluding.Beer + Eggs +
##     Fish..Seafood + Fruits...Excluding.Wine + Meat + Milk...Excluding.Butter +
##     Offals + Oilcrops + Pulses + Spices + Starchy.Roots + Stimulants +
##     Sugar.Crops + Sugar...Sweeteners + Treenuts + Vegetal.Products +
##     Vegetable.Oils + Vegetables + Miscellaneous + Obesity + Undernourished +
##     Population) - Undernourished - Population
## Model 2: Recovered^0.5 ~ Alcoholic.Beverages + Animal.Products + Animal.fats +
##     Aquatic.Products..Other + Cereals...Excluding.Beer + Eggs +
##     Fish..Seafood + Fruits...Excluding.Wine + Meat + Milk...Excluding.Butter +
##     Offals + Oilcrops + Pulses + Spices + Starchy.Roots + Stimulants +
##     Sugar.Crops + Sugar...Sweeteners + Treenuts + Vegetal.Products +
##     Vegetable.Oils + Vegetables + Miscellaneous + Obesity + Undernourished +
##     Population
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    129 48.606
## 2    127 48.007  2   0.59912 0.7925  0.455
```

The full model is better than the model with fewer variables

```
## [1] 1.460701
```

The training error is 1.640541

```
## [1] 2.393609
```

The prediction error is 2.058205

PCA Regression

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.6706 1.52258 1.38578 1.32461 1.23362 1.19792 1.07838
## Proportion of Variance 0.2641 0.08586 0.07112 0.06498 0.05636 0.05315 0.04307
## Cumulative Proportion  0.2641 0.35001 0.42114 0.48612 0.54249 0.59563 0.63870
##                            PC8    PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.04809 1.0340 0.94730 0.91692 0.87057 0.84054 0.80807
## Proportion of Variance 0.04068 0.0396 0.03324 0.03114 0.02807 0.02617 0.02418
## Cumulative Proportion  0.67939 0.7190 0.75222 0.78336 0.81143 0.83760 0.86178
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.78250 0.76446 0.67024 0.64173 0.61731 0.59675 0.56060
## Proportion of Variance 0.02268 0.02164 0.01664 0.01525 0.01411 0.01319 0.01164
## Cumulative Proportion  0.88446 0.90611 0.92274 0.93800 0.95211 0.96530 0.97694
##                           PC22    PC23    PC24     PC25      PC26      PC27
## Standard deviation     0.51305 0.44538 0.40129 0.000571 0.0004148 7.522e-06
## Proportion of Variance 0.00975 0.00735 0.00596 0.000000 0.0000000 0.000e+00
## Cumulative Proportion  0.98669 0.99404 1.00000 1.000000 1.0000000 1.000e+00
```

Will use the first 19 components in the regression as they account for 95% of the variance

```
## Warning: package 'pls' was built under R version 3.6.2
```

```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
## Data:    X dimension: 92 26
##  Y dimension: 92 1
## Fit method: svdpc
## Number of components considered: 19
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X          100.0000  100.000   100.00   100.00   100.00   100.00   100.00
## Recovered    0.7948    9.881    27.19    27.19    28.97    31.17    31.41
##            8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X           100.00   100.00    100.00    100.00     100.0    100.00    100.00
## Recovered    31.41    31.45     33.86     35.52      36.3     37.91     38.07
##            15 comps  16 comps  17 comps  18 comps  19 comps
## X            100.00    100.00    100.00    100.00    100.00
## Recovered     38.46     40.14     40.51     40.93     40.99
```

In training 40% of the variance is explained.

```
## [1] 1.406283
```

The training error is 1.608312

```
## [1] 2.169666
```

The prediction error is 1.835051

Main Conclusion

- The amount of undernourished, obese, and population were not significant variables.
- OLS regression is not a good model with a low r squared and a high prediction error
- PCA regression is a much better regression model with a much smaller prediction error