Factor Analysis

Russell and Frances

September 2022

Contents

Factor analysis on health variables, with null hypothesis that two factors are sufficient. P-value <0.01 so reject null and conclude 2 factors are not sufficient.

```
factanal(df1, factors = 2)
```

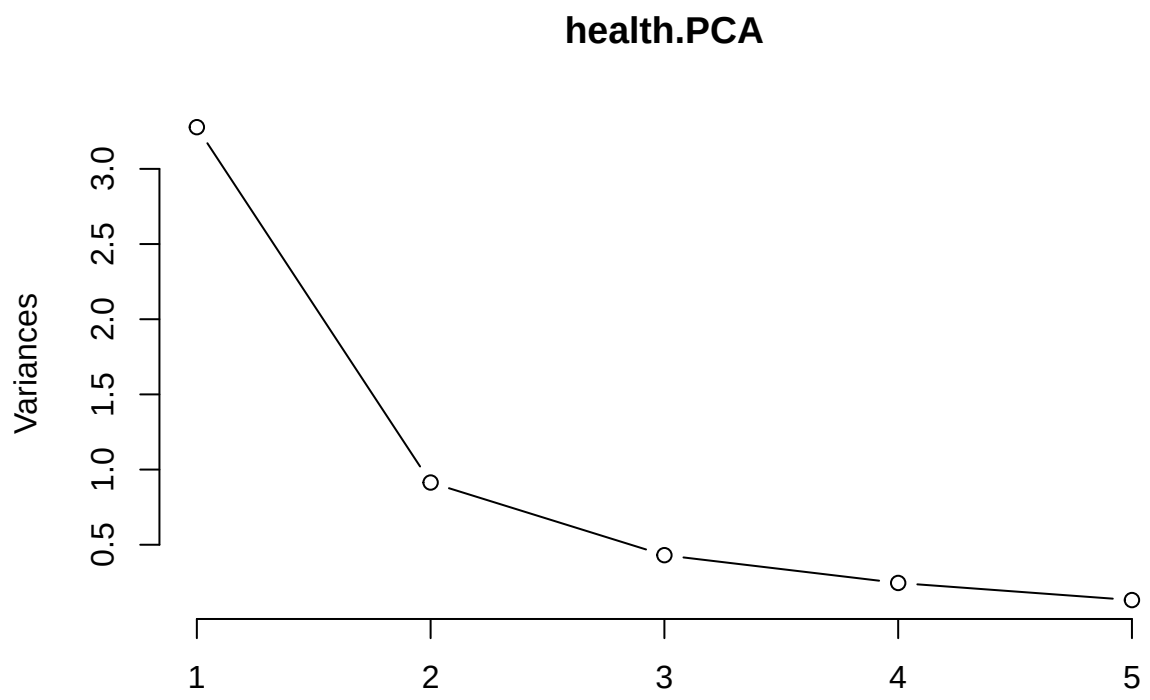
```
##
## Call:
## factanal(x = df1, factors = 2)
##
## Uniquenesses:
## SECCHI   NH4N   CHLA     TP     TN
##  0.353  0.408  0.005  0.419  0.005
##
## Loadings:
##        Factor1 Factor2
## SECCHI -0.632  -0.497
## NH4N    0.764
## CHLA    0.149   0.986
## TP      0.412   0.641
## TN      0.902   0.426
##
##                Factor1 Factor2
## SS loadings      1.988   1.821
## Proportion Var   0.398   0.364
## Cumulative Var   0.398   0.762
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 65.57 on 1 degree of freedom.
## The p-value is 5.59e-16
```

```
health.PCA <- prcomp(df1, center=TRUE, scale=TRUE)
health.PCA
```

```
## Standard deviations (1, .., p=5):
## [1] 1.8103459 0.9559851 0.6563187 0.4958903 0.3634268
##
## Rotation (n x k) = (5 x 5):
##               PC1         PC2         PC3         PC4         PC5
```

```
## SECCHI  0.4718654 -0.05415131  0.71443627 -0.33614772  0.38858237
## NH4N   -0.3804385  0.68273926  0.32640483 -0.38001162 -0.37173198
## CHLA   -0.4160810 -0.58414236 -0.08611696 -0.69136513 -0.01588745
## TP     -0.4448543 -0.36615759  0.61121174  0.50549627 -0.19729917
## TN     -0.5114595  0.23588357 -0.04521927  0.09530825  0.81953627
```

```
plot(health.PCA, type="l")
```




```
S <- cov(df1)
S
```

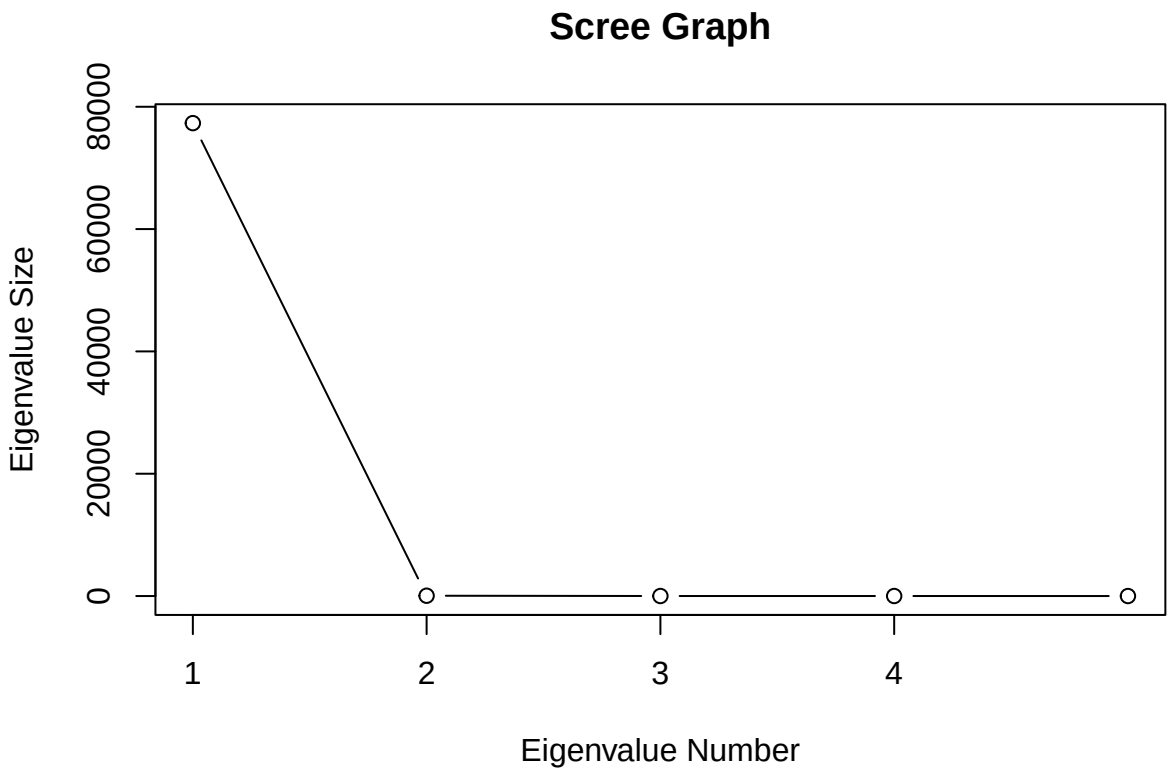
```
##               SECCHI          NH4N          CHLA            TP           TN
## SECCHI  5.086583e+00 -0.0078525794  -3.701782370  -11.00491286  -490.520745
## NH4N   -7.852579e-03  0.0000467284   0.003983157    0.02341271     1.386062
## CHLA   -3.701782e+00  0.0039831570   7.879622903   17.81020624   433.309014
## TP     -1.100491e+01  0.0234127142  17.810206240   83.60680467  1640.359387
## TN     -4.905207e+02  1.3860622907 433.309013568 1640.35938689 77280.925931
```

```
S.eigen <- eigen(S)
S.eigen
```

```
## eigen() decomposition
## $values
## [1] 7.732131e+04 5.042512e+01 4.115416e+00 1.646826e+00 1.919020e-05
```

```
##
## $vectors
##               [,1]          [,2]          [,3]          [,4]          [,5]
## [1,]  6.345977e-03  0.0155691968  0.3567397498  9.340525e-01  1.593886e-04
## [2,] -1.792868e-05  0.0001317947  0.0006607602 -8.379513e-05 -9.999998e-01
## [3,] -5.608307e-03 -0.1882004126 -0.9164135927  3.531782e-01 -6.598277e-04
## [4,] -2.123445e-02 -0.9817605977  0.1814106493 -5.277692e-02 -4.718796e-06
## [5,] -9.997387e-01  0.0220071858  0.0035521515  5.068759e-03  2.274682e-05
```

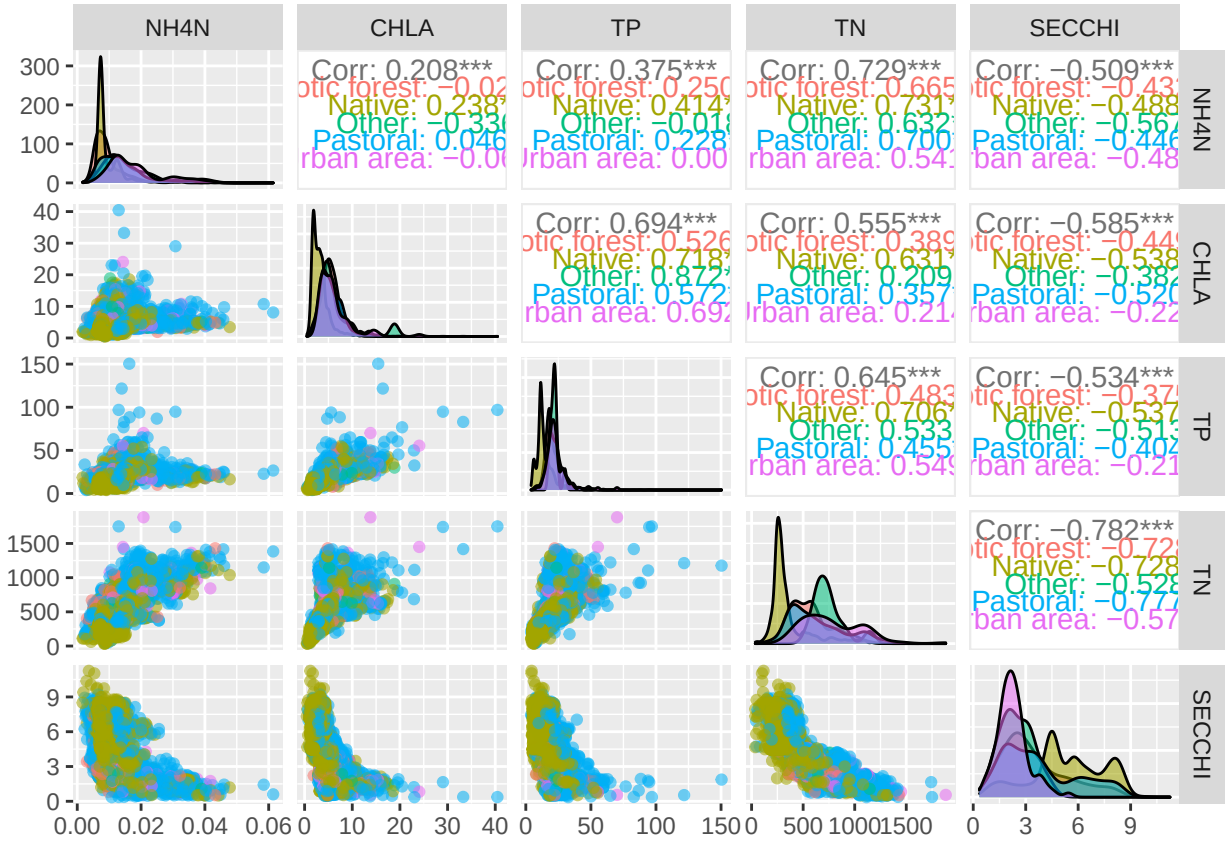
```
plot(S.eigen$values, xlab = 'Eigenvalue Number', ylab = 'Eigenvalue Size', main = 'Scree Graph', type =
axis(1, at = seq(1, 4, by = 1))
```




```
C <- as.matrix(S.eigen$vectors[,1:2])

D <- matrix(0, dim(C)[2], dim(C)[2])
diag(D) <- S.eigen$values[1:2]
S.loadings <- C %*% sqrt(D)
S.loadings
```

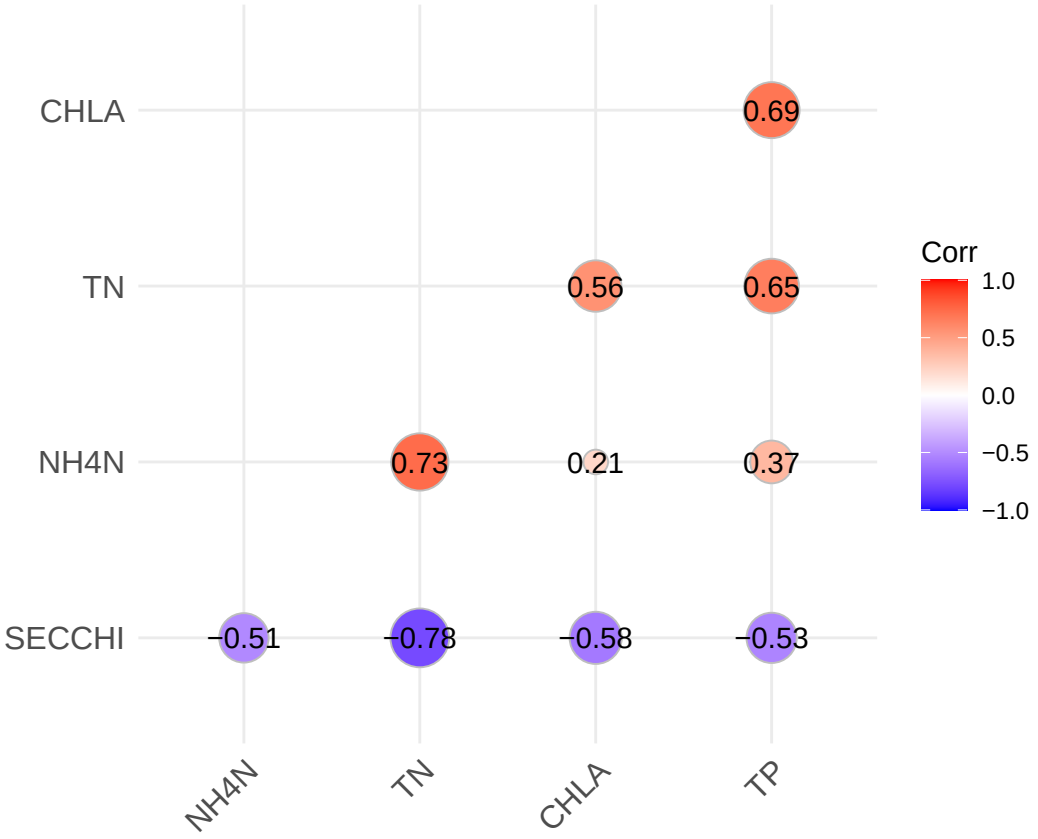
```
##               [,1]         [,2]
## [1,]  1.764607e+00  0.110557873
## [2,] -4.985377e-03  0.000935883
## [3,] -1.559486e+00 -1.336423302
## [4,] -5.904603e+00 -6.971545501
## [5,] -2.779944e+02  0.156274450
```

```
ggpairs(df2, columns=2:6, ggplot2::aes(colour=land, alpha=.7))
```

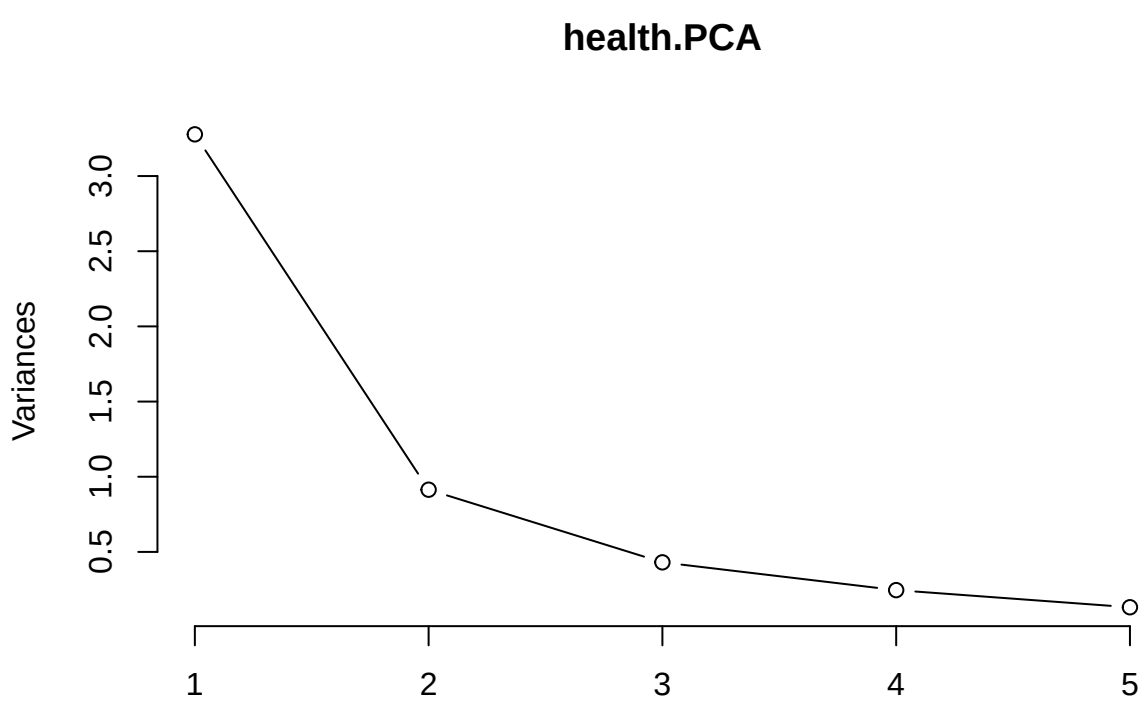



```
ggcorrplot(cor(df1),
method = "circle",
hc.order = TRUE,
type = "lower",
lab = TRUE)
```

```
## Warning: ‘guides(<scale> = FALSE)‘ is deprecated. Please use ‘guides(<scale> =
## "none")‘ instead.
```




```
health.PCA <- prcomp(df1, center=TRUE, scale=TRUE)
plot(health.PCA, type="l")
```




```
summary(health.PCA)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5
## Standard deviation     1.8103 0.9560 0.65632 0.49589 0.36343
## Proportion of Variance 0.6555 0.1828 0.08615 0.04918 0.02642
## Cumulative Proportion  0.6555 0.8383 0.92440 0.97358 1.00000
```

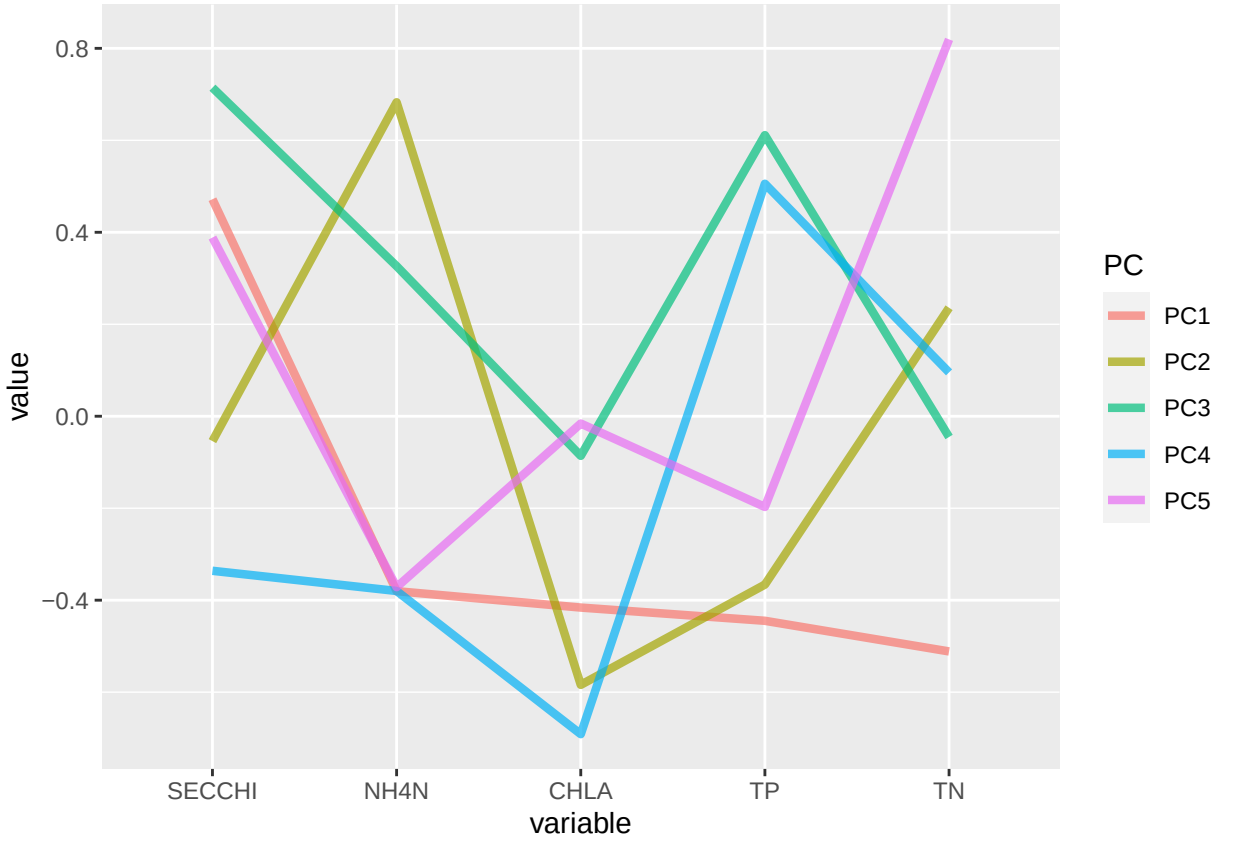
1st PC captures 65.6% of variance so we continue with just PC1

eigenvectors

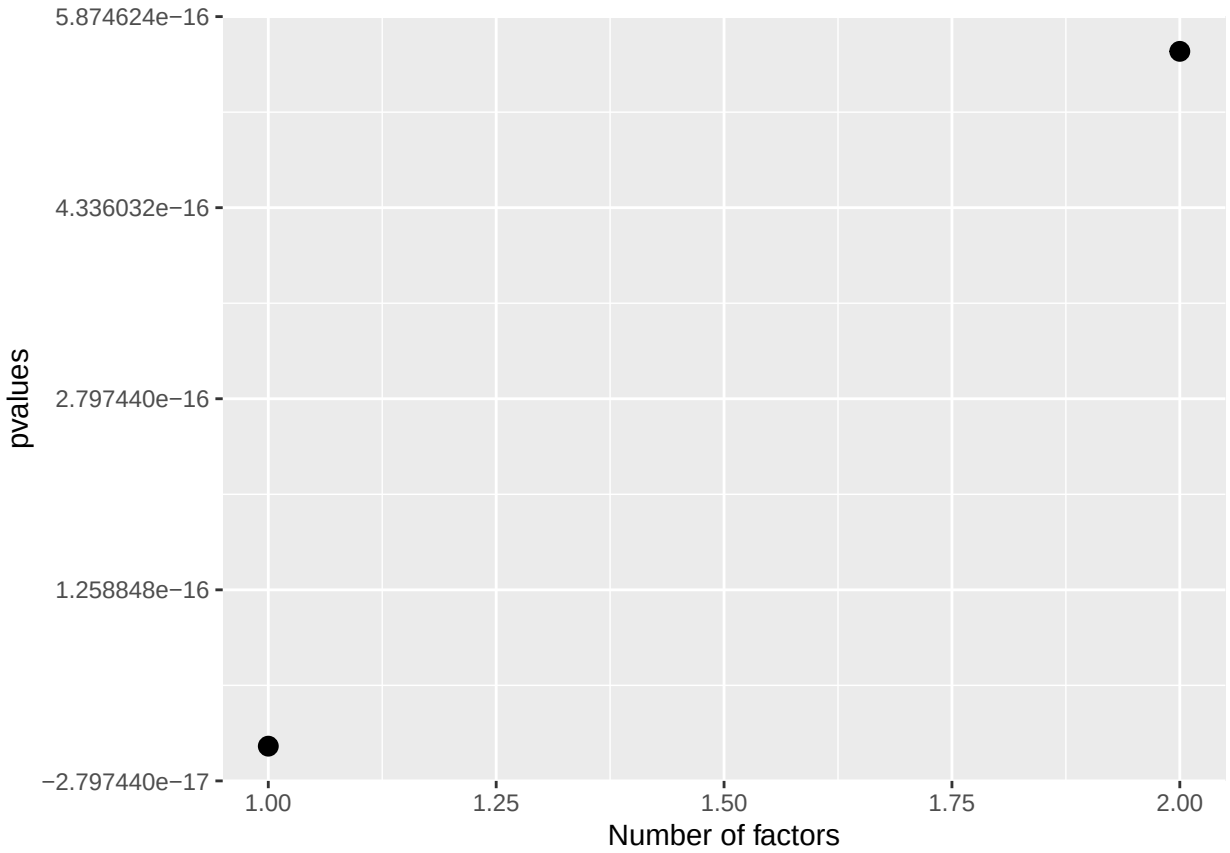
```
health.PCA$rotation
```

```
##               PC1         PC2         PC3         PC4         PC5
## SECCHI  0.4718654 -0.05415131  0.71443627 -0.33614772  0.38858237
## NH4N   -0.3804385  0.68273926  0.32640483 -0.38001162 -0.37173198
## CHLA   -0.4160810 -0.58414236 -0.08611696 -0.69136513 -0.01588745
## TP     -0.4448543 -0.36615759  0.61121174  0.50549627 -0.19729917
## TN     -0.5114595  0.23588357 -0.04521927  0.09530825  0.81953627
```

```
Rotation.df <- data.frame(t(health.PCA$rotation))
Rotation.df$PC <- c("PC1", "PC2", "PC3", "PC4", "PC5")
Rotation.melt <- melt(Rotation.df, id.vars = "PC")
ggplot(Rotation.melt) +
geom_line(aes(x=variable, y=value, group=PC, col=PC), size=1.5, alpha=.7)
```




```
pvalues <- rep(0, 2)
for(f in 1:2)
  pvalues[f] <- factanal(df1, factors = f)$PVAL
ggplot(data.frame(pvalues), aes(x=1:2, y=pvalues)) +
geom_point(size=3) +
xlab("Number of factors")
```



Factor analysis on health variables, with null hypothesis that two factors are sufficient. P-value <0.01 so reject null and conclude 2 factors are not sufficient.

```
factanal(df1, factors = 2)
```

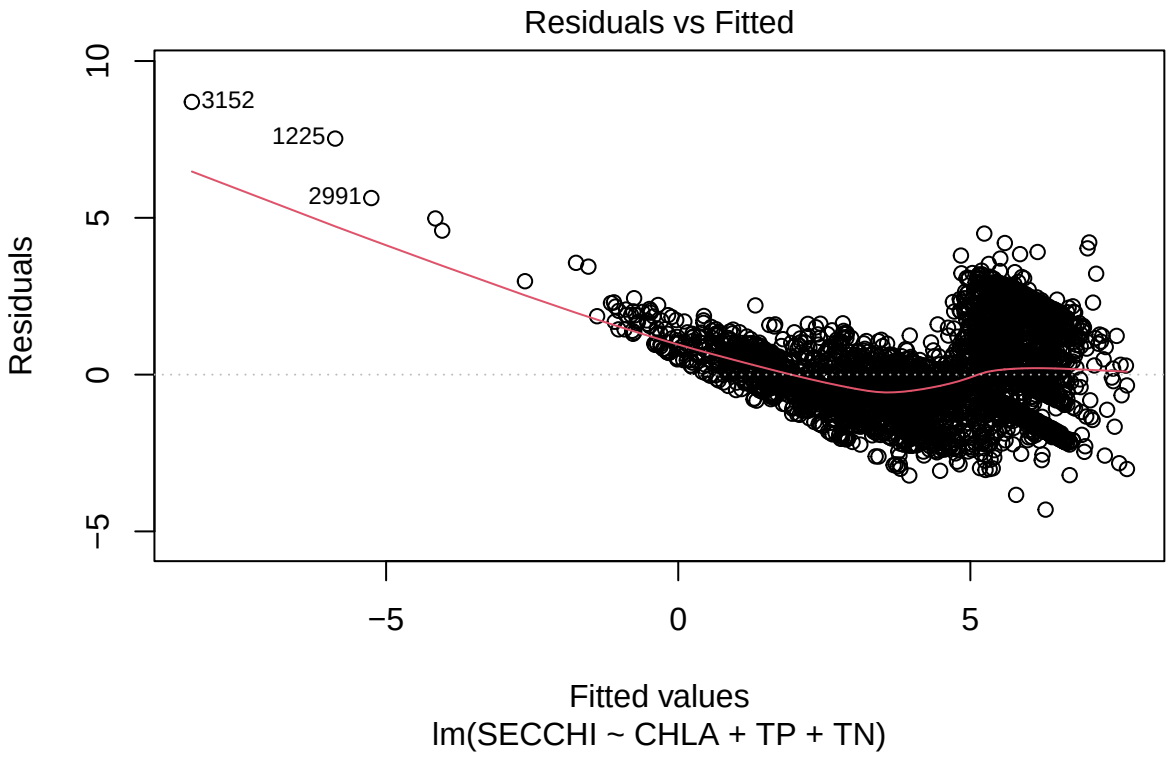
```
##
## Call:
## factanal(x = df1, factors = 2)
##
## Uniquenesses:
## SECCHI   NH4N   CHLA     TP     TN
##  0.353  0.408  0.005  0.419  0.005
##
## Loadings:
##        Factor1 Factor2
## SECCHI -0.632  -0.497
## NH4N    0.764
## CHLA    0.149   0.986
## TP      0.412   0.641
## TN      0.902   0.426
##
##                Factor1 Factor2
## SS loadings      1.988   1.821
## Proportion Var   0.398   0.364
## Cumulative Var   0.398   0.762
##
```

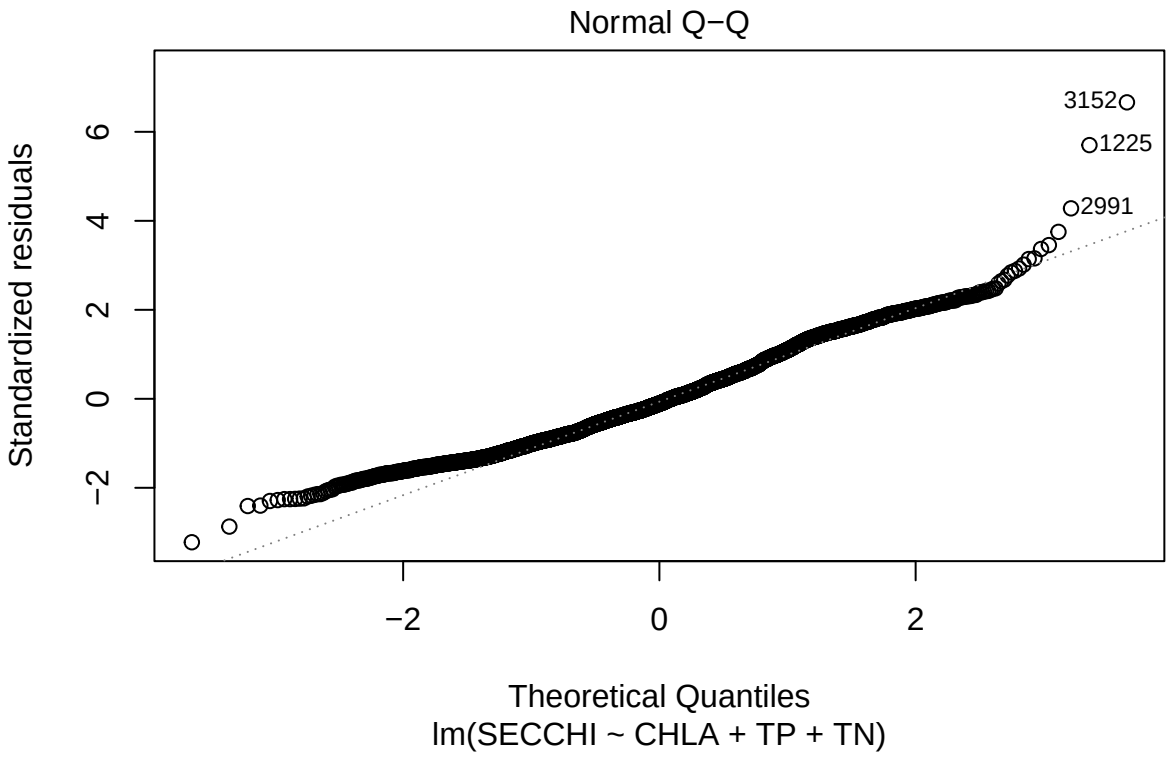
```
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 65.57 on 1 degree of freedom.
## The p-value is 5.59e-16
```

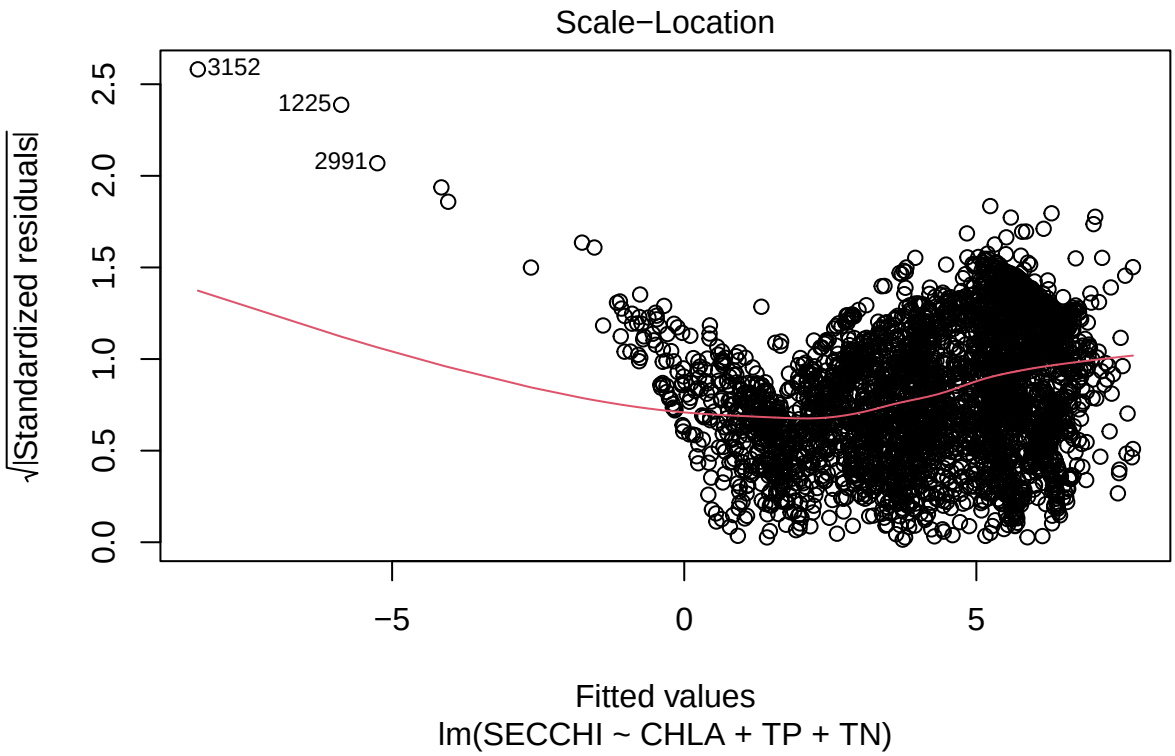
```
full.lm <- lm(data=df1, SECCHI~CHLA+TP+TN)
summary(full.lm)
```

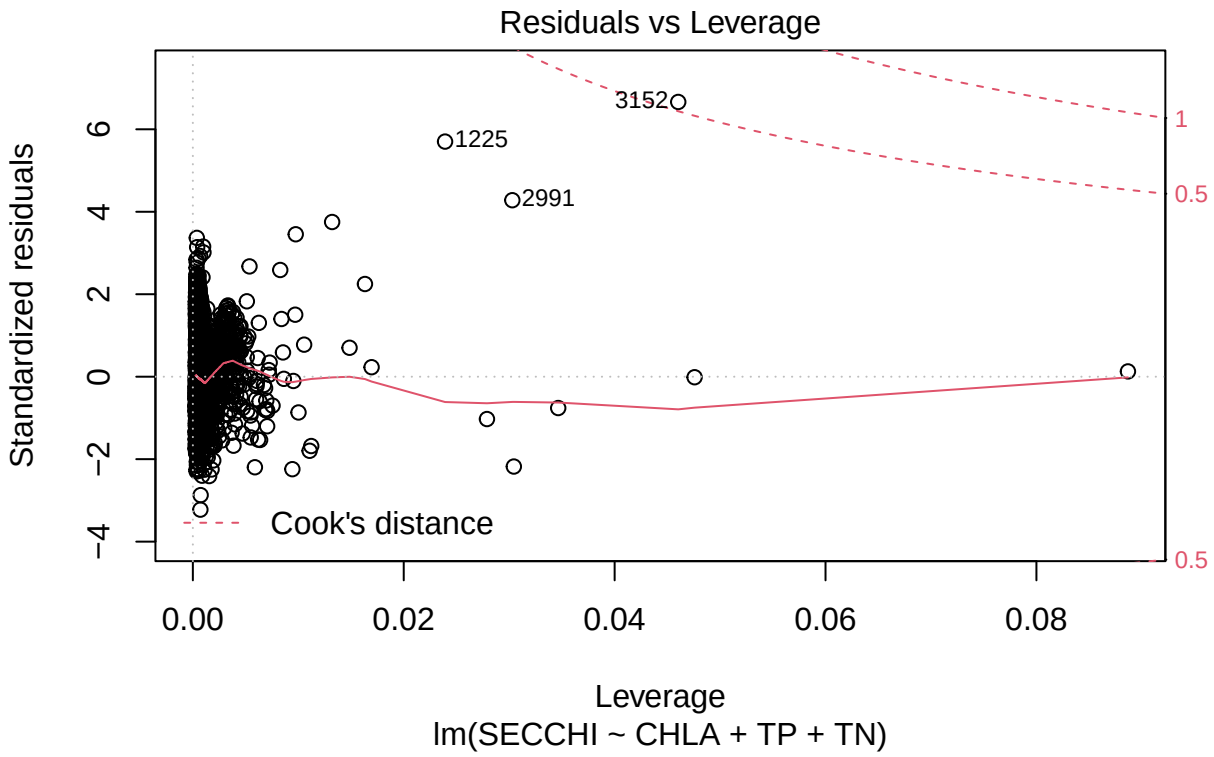
```
##
## Call:
## lm(formula = SECCHI ~ CHLA + TP + TN, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3054 -1.0322 -0.1452  0.8609  8.6942
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.837206   0.051513 152.142  < 2e-16 ***
## CHLA        -0.215479   0.010932 -19.712  < 2e-16 ***
## TP           0.025882   0.003654   7.084 1.66e-12 ***
## TN          -0.005688   0.000104 -54.671  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.336 on 3798 degrees of freedom
## Multiple R-squared:  0.6494, Adjusted R-squared:  0.6491
## F-statistic:  2345 on 3 and 3798 DF,  p-value: < 2.2e-16
```

```
plot(full.lm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SECCHI ~ CHLA + TP + TN)
3152
1225
2991
10
5
0
−5
−5
0
5


Normal Q−Q
Standardized residuals
−2
0
2
4
6
3152
1225
2991
−2
0
2
Theoretical Quantiles
lm(SECCHI ~ CHLA + TP + TN)

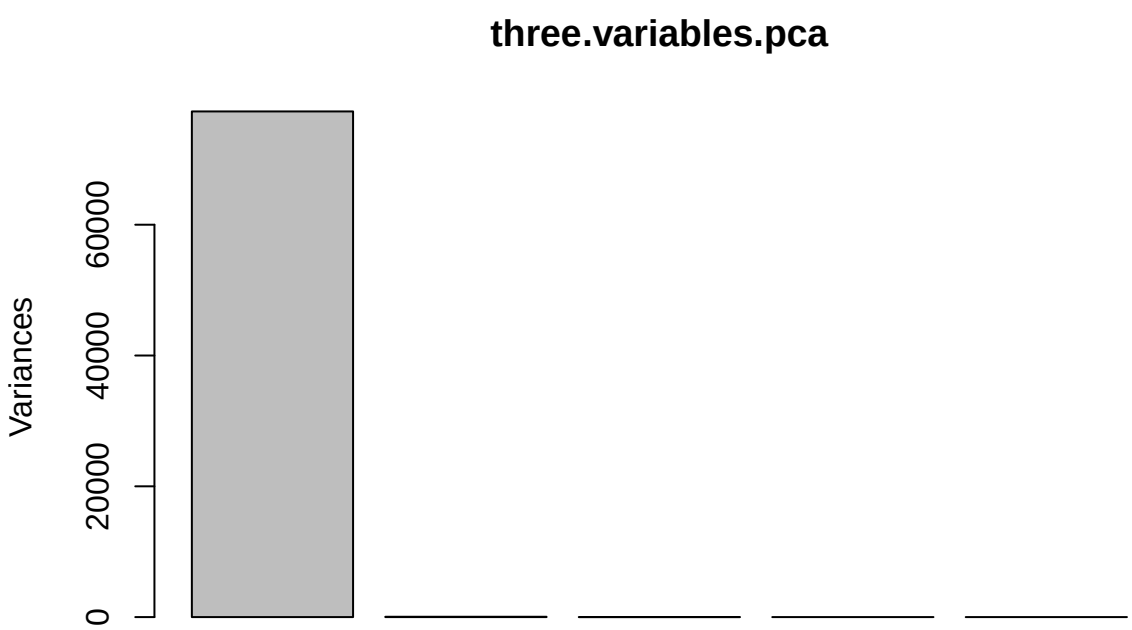

Scale−Location
3152
1225
2991
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
−5
0
5
Fitted values
lm(SECCHI ~ CHLA + TP + TN)




```
three.variables.pca <- prcomp(df1)
summary(three.variables.pca)
```

```
## Importance of components:
##                             PC1     PC2     PC3     PC4      PC5
## Standard deviation     278.0671 7.10106 2.02865 1.28329 0.004381
## Proportion of Variance   0.9993 0.00065 0.00005 0.00002 0.000000
## Cumulative Proportion    0.9993 0.99993 0.99998 1.00000 1.000000
```

```
plot(three.variables.pca)
```




```
three.variables.pca$rotation
```

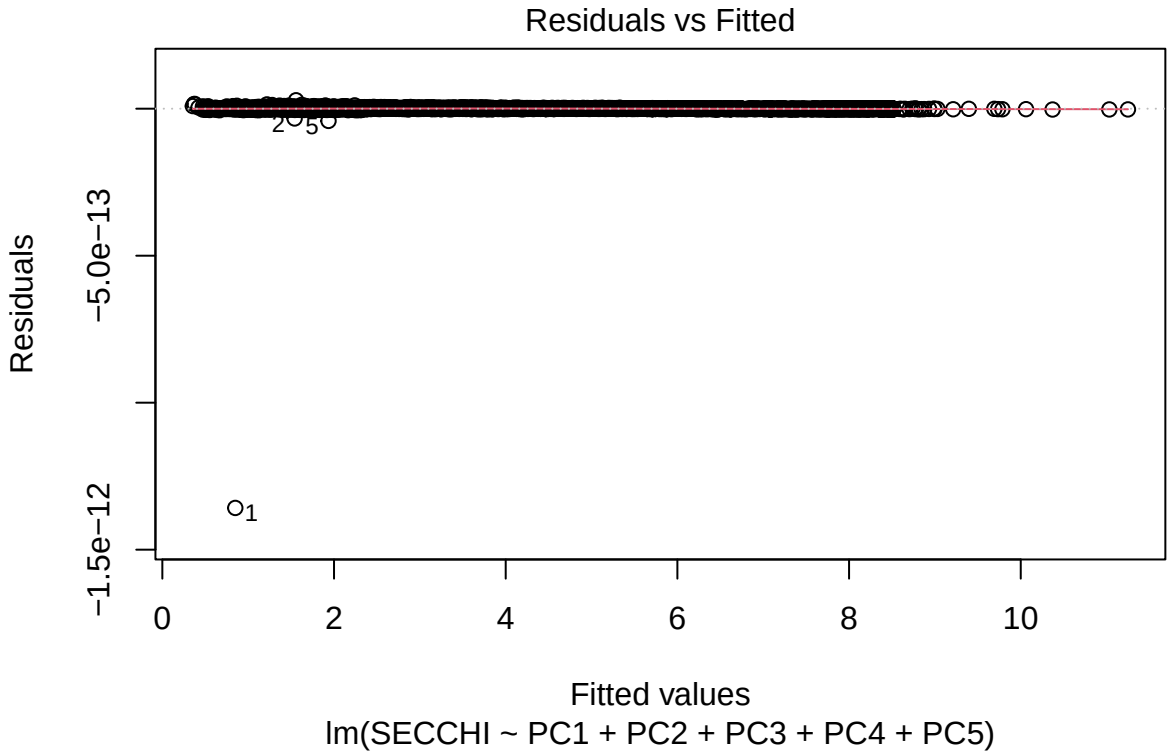
```
##                  PC1           PC2           PC3           PC4           PC5
## SECCHI  6.345977e-03 -0.0155691968  0.3567397498  9.340525e-01  1.593886e-04
## NH4N   -1.792868e-05 -0.0001317947  0.0006607602 -8.379513e-05 -9.999998e-01
## CHLA   -5.608307e-03  0.1882004126 -0.9164135927  3.531782e-01 -6.598277e-04
## TP     -2.123445e-02  0.9817605977  0.1814106493 -5.277692e-02 -4.718796e-06
## TN     -9.997387e-01 -0.0220071858  0.0035521515  5.068759e-03  2.274682e-05
```

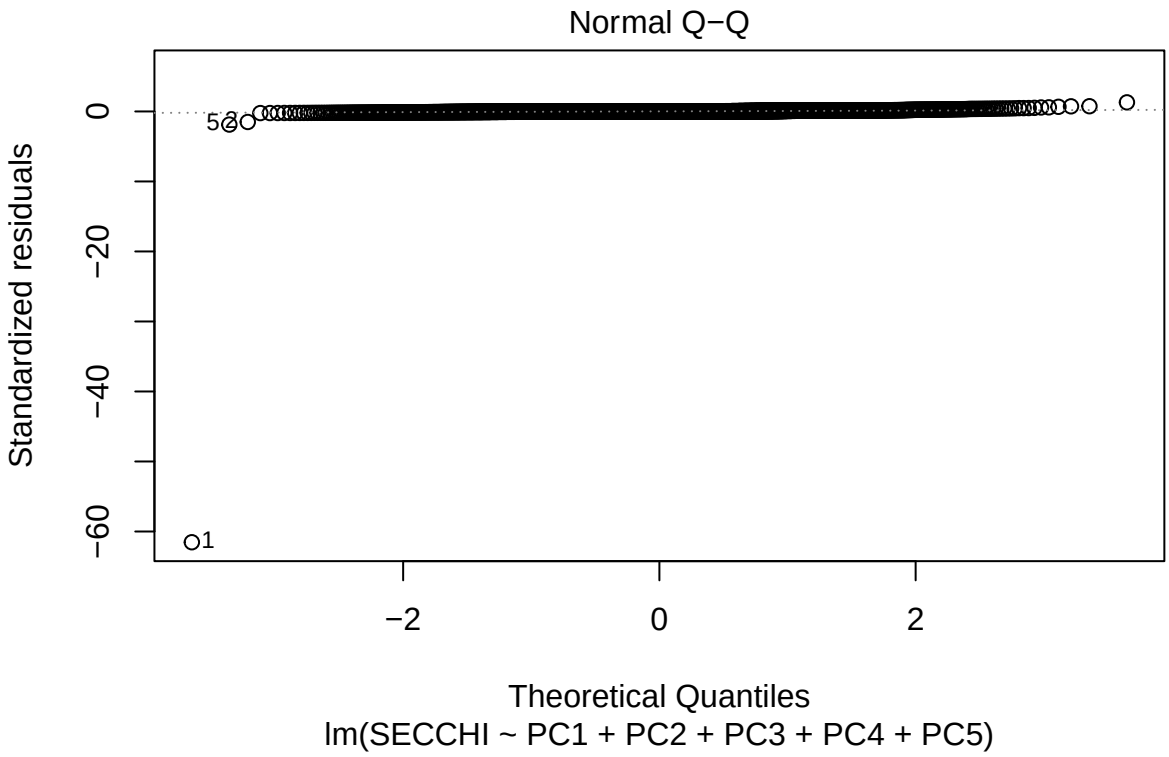
```
three.variables.pca.x <- three.variables.pca$x
df1$PC1 <- three.variables.pca.x[,1]
df1$PC2 <- three.variables.pca.x[,2]
df1$PC3 <- three.variables.pca.x[,3]
df1$PC4 <- three.variables.pca.x[,4]
df1$PC5 <- three.variables.pca.x[,5]

full.lm.pca <- lm(data=df1, SECCHI~PC1+PC2+PC3+PC4+PC5)
summary(full.lm.pca)
```

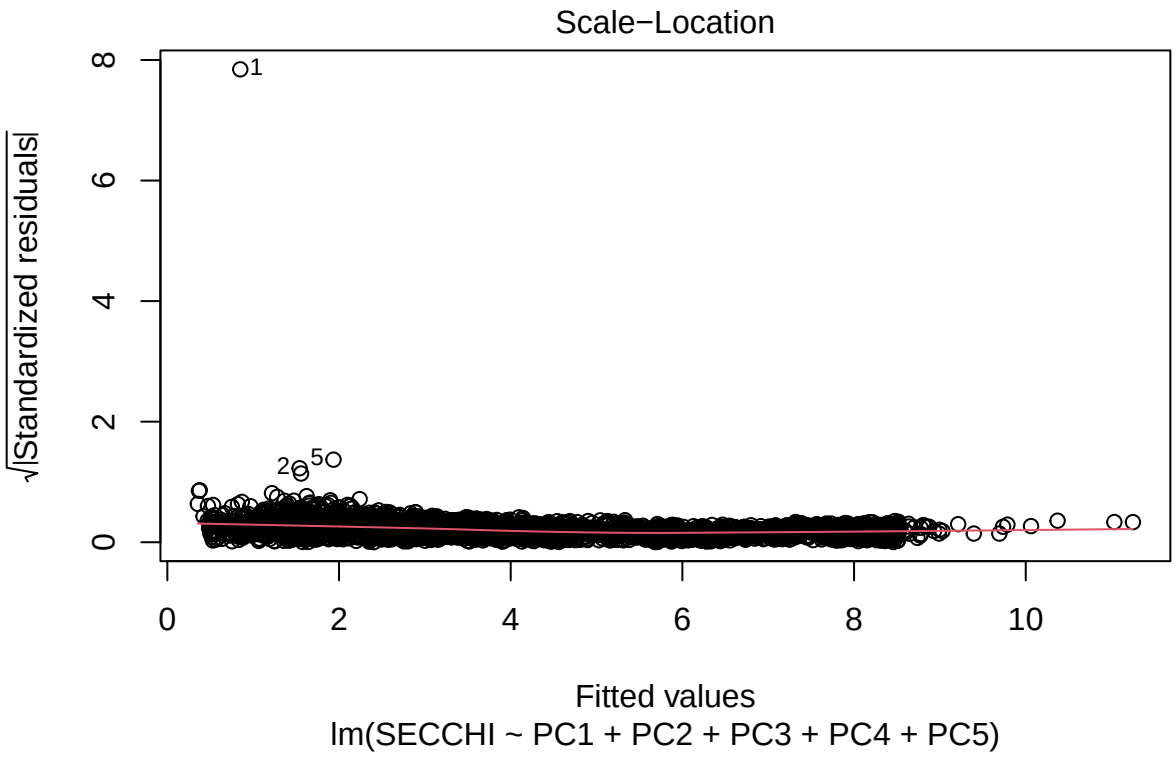
```
##
## Call:
## lm(formula = SECCHI ~ PC1 + PC2 + PC3 + PC4 + PC5, data = df1)
##
## Residuals:
```

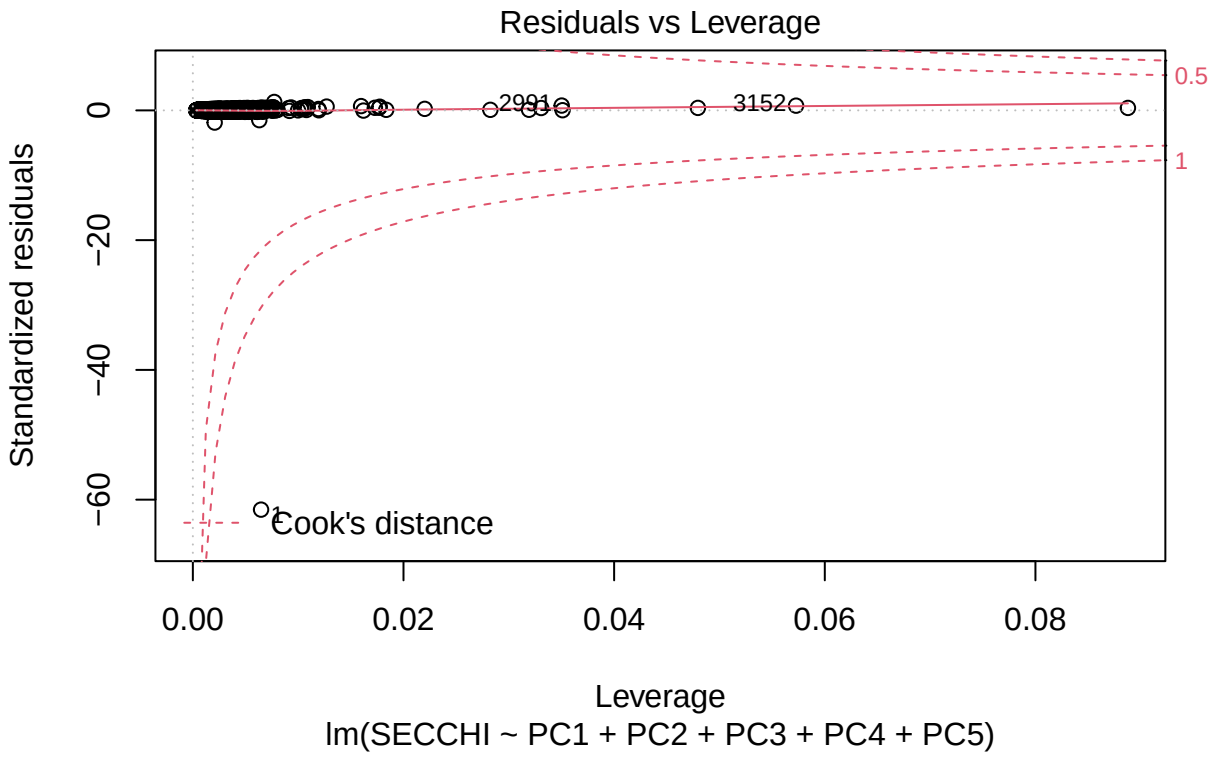
```
##        Min         1Q     Median         3Q        Max
## -1.358e-12 -6.000e-16  1.200e-16  1.050e-15  2.865e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  4.451e+00  3.591e-16  1.240e+16   <2e-16 ***
## PC1          6.346e-03  1.291e-18  4.914e+15   <2e-16 ***
## PC2         -1.557e-02  5.057e-17 -3.079e+14   <2e-16 ***
## PC3          3.567e-01  1.770e-16  2.015e+15   <2e-16 ***
## PC4          9.341e-01  2.798e-16  3.338e+15   <2e-16 ***
## PC5          1.594e-04  8.198e-14  1.944e+09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.214e-14 on 3796 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.888e+30 on 5 and 3796 DF,  p-value: < 2.2e-16
```

```
plot(full.lm.pca)
```




Normal Q−Q
Standardized residuals
0
−20
−40
−60
5
2
1
−2
0
2
Theoretical Quantiles
lm(SECCHI ~ PC1 + PC2 + PC3 + PC4 + PC5)

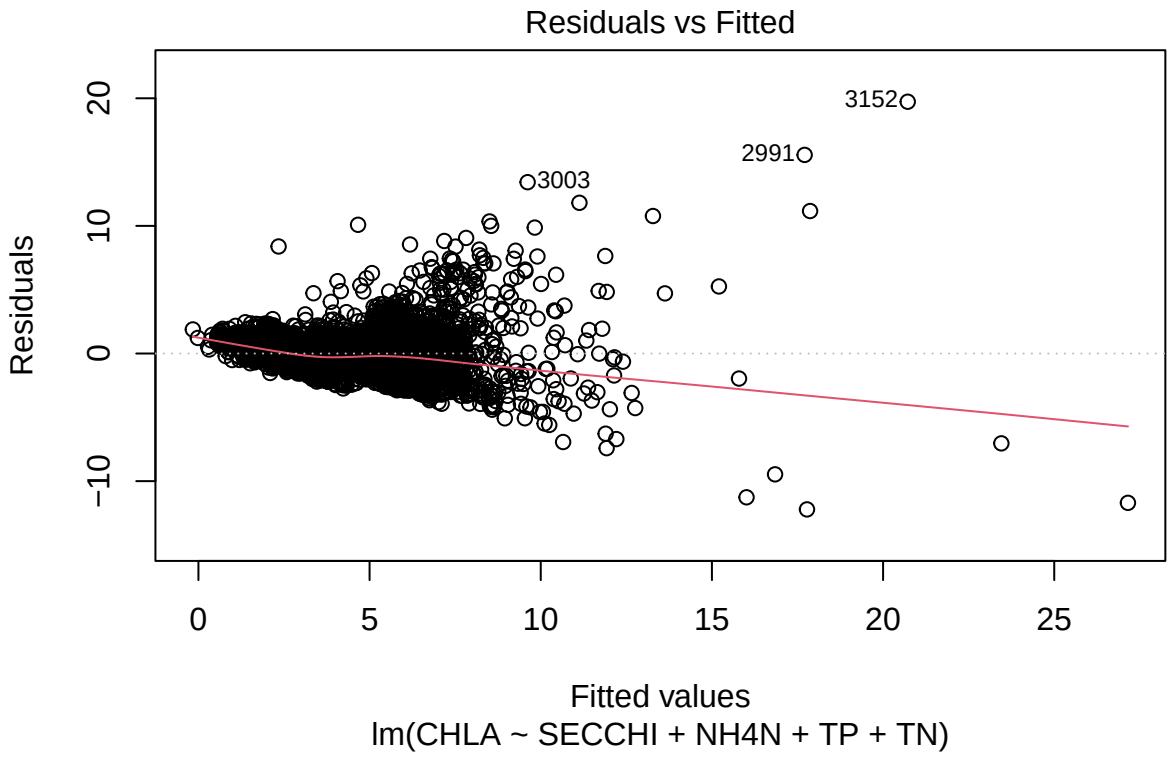

Scale−Location
√|Standardized residuals|
0
2
4
6
8
0
2
4
6
8
10
1
2
5
Fitted values
lm(SECCHI ~ PC1 + PC2 + PC3 + PC4 + PC5)

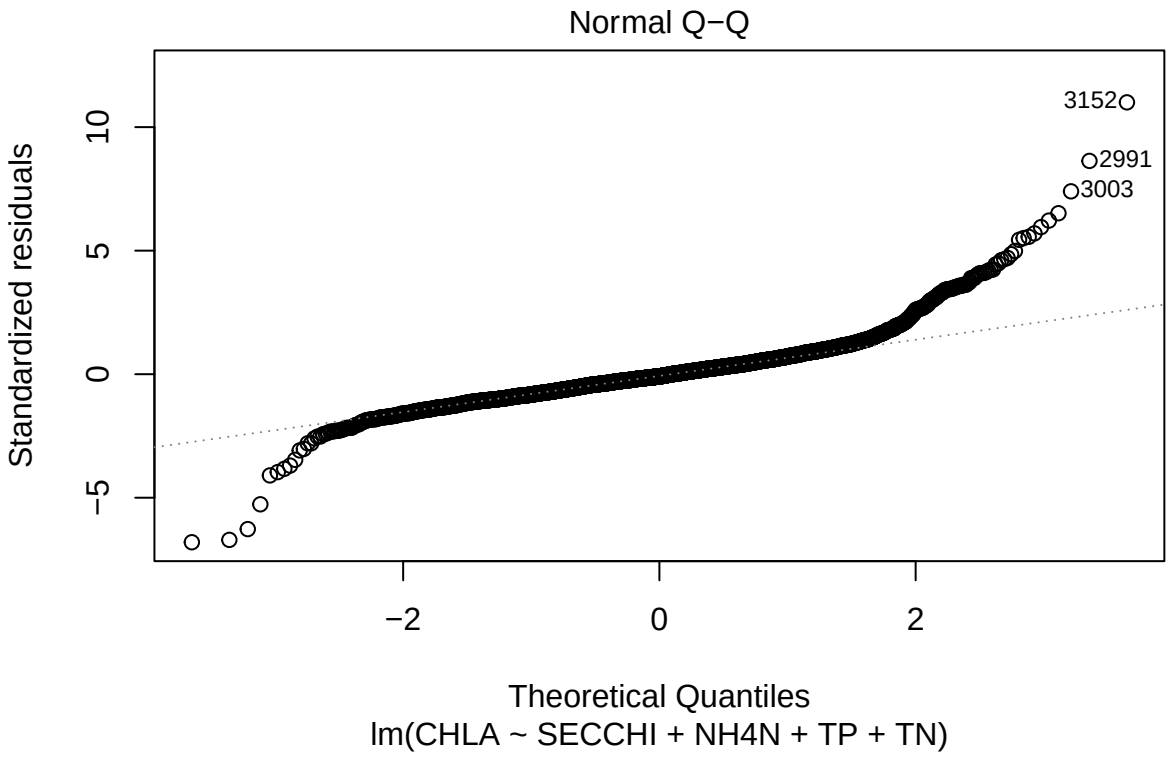


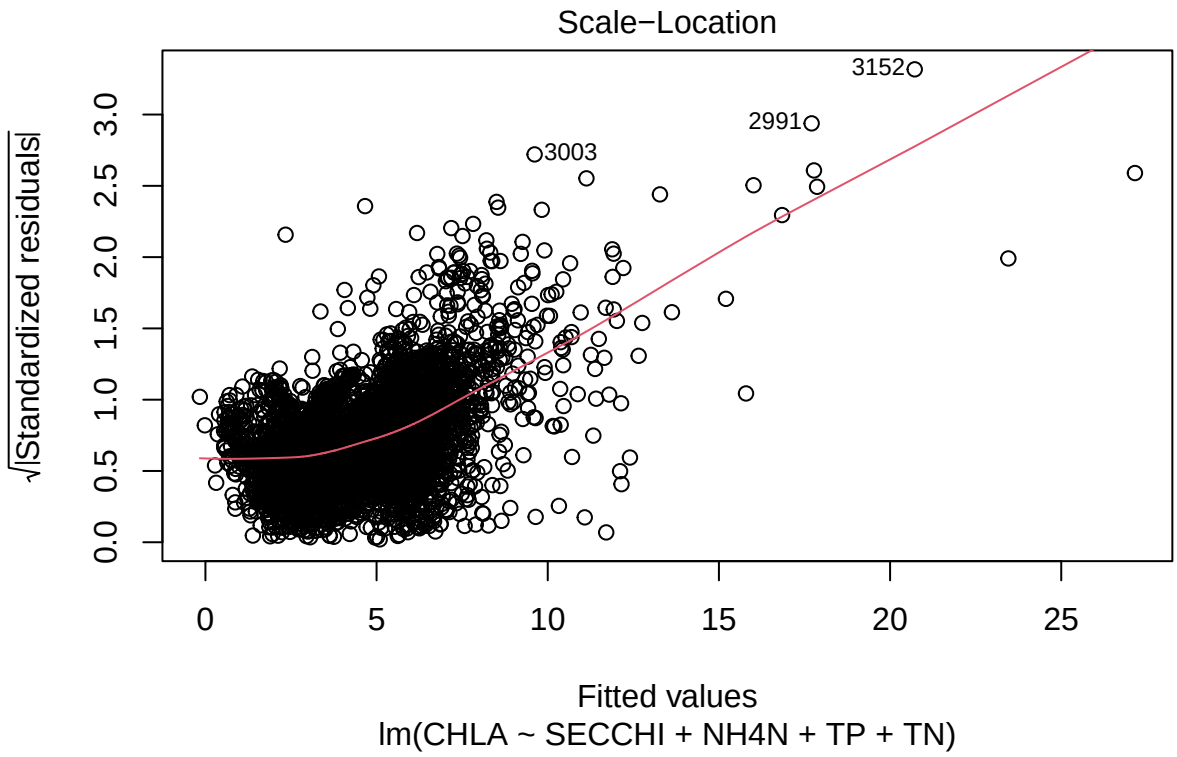

```
full.lm <- lm(data=df1, CHLA~SECCHI+NH4N+TP+TN)
summary(full.lm)
```

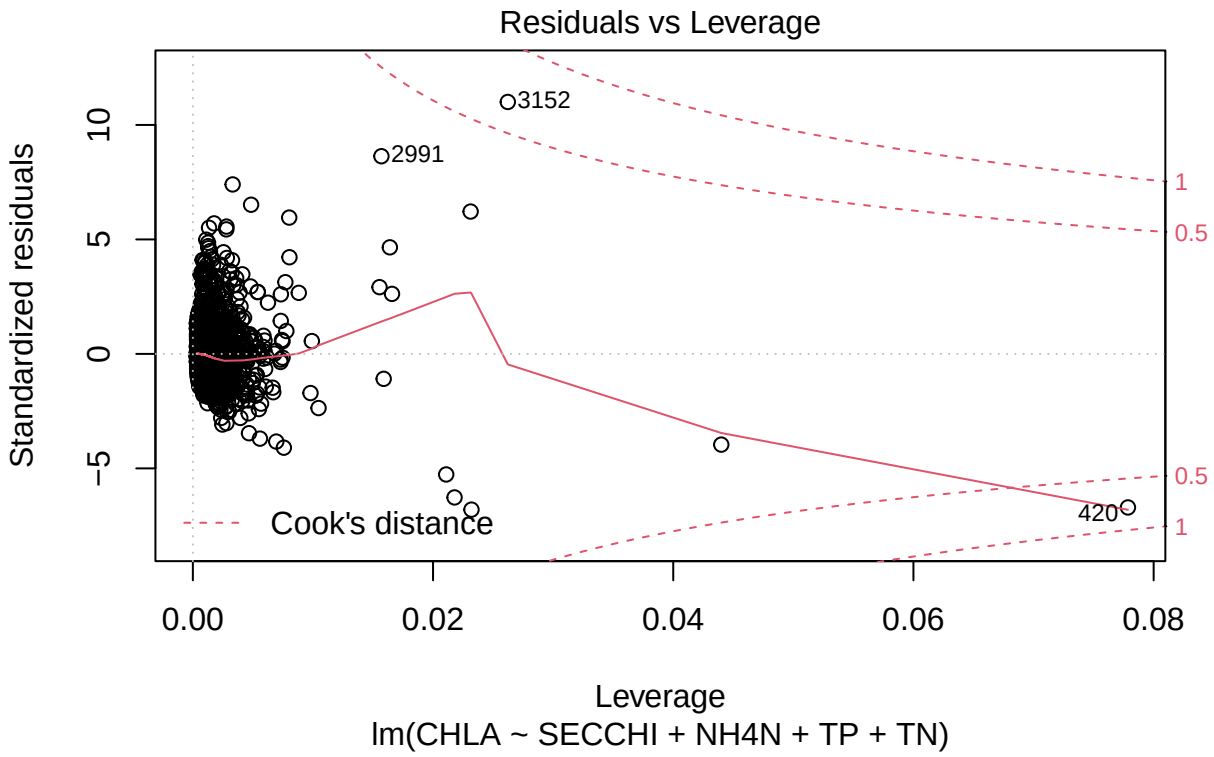
```
##
## Call:
## lm(formula = CHLA ~ SECCHI + NH4N + TP + TN, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2152  -1.0248  -0.1623   0.7707  19.7291
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.753e+00  1.774e-01  21.152  < 2e-16 ***
## SECCHI      -3.804e-01  2.122e-02 -17.931  < 2e-16 ***
## NH4N        -1.134e+02  6.471e+00 -17.524  < 2e-16 ***
## TP           1.579e-01  4.295e-03  36.771  < 2e-16 ***
## TN           1.874e-03  2.414e-04   7.762 1.07e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.817 on 3797 degrees of freedom
## Multiple R-squared:  0.5814, Adjusted R-squared:  0.581
## F-statistic:  1319 on 4 and 3797 DF,  p-value: < 2.2e-16
```

```
plot(full.lm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(CHLA ~ SECCHI + NH4N + TP + TN)
3152
2991
3003
20
10
0
−10
0
5
10
15
20
25


Normal Q−Q
Standardized residuals
10
5
0
−5
3152
2991
3003
−2
0
2
Theoretical Quantiles
lm(CHLA ~ SECCHI + NH4N + TP + TN)

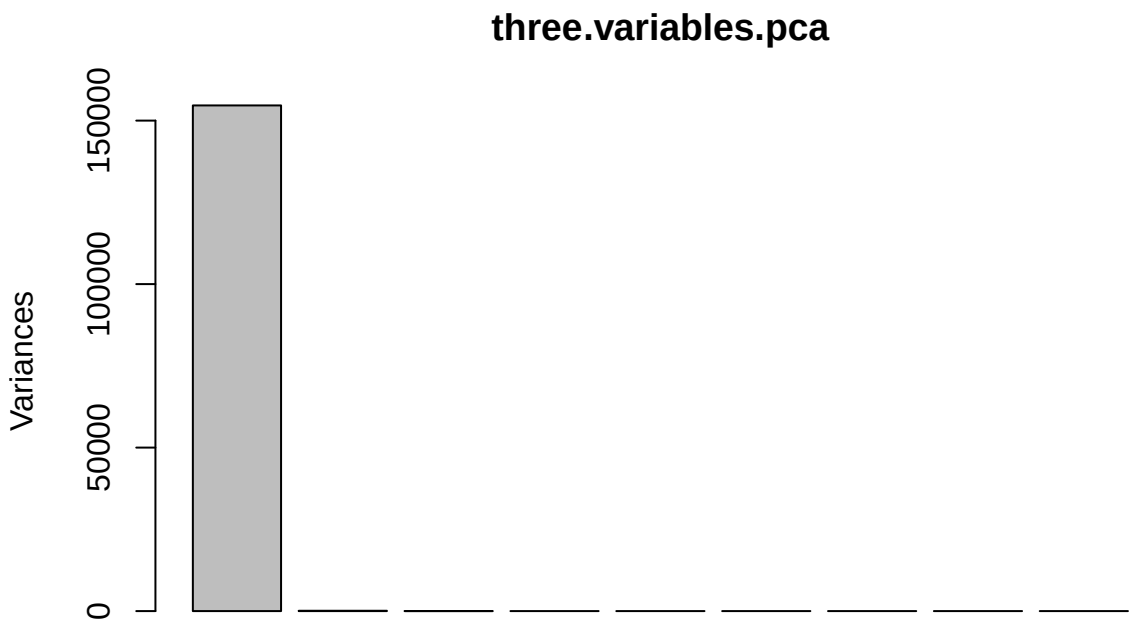

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3152
2991
3003
0
5
10
15
20
25
Fitted values
lm(CHLA ~ SECCHI + NH4N + TP + TN)




```
three.variables.pca <- prcomp(df1[,-2])
summary(three.variables.pca)
```

```
## Importance of components:
##                             PC1      PC2     PC3     PC4      PC5       PC6
## Standard deviation     393.2463 10.04242 2.86894 1.81484 0.004381 2.968e-13
## Proportion of Variance   0.9993  0.00065 0.00005 0.00002 0.000000 0.000e+00
## Cumulative Proportion    0.9993  0.99993 0.99998 1.00000 1.000000 1.000e+00
##                              PC7       PC8       PC9
## Standard deviation     1.068e-14 4.976e-15 3.343e-15
## Proportion of Variance 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00
```

```
plot(three.variables.pca)
```




```
three.variables.pca$rotation
```

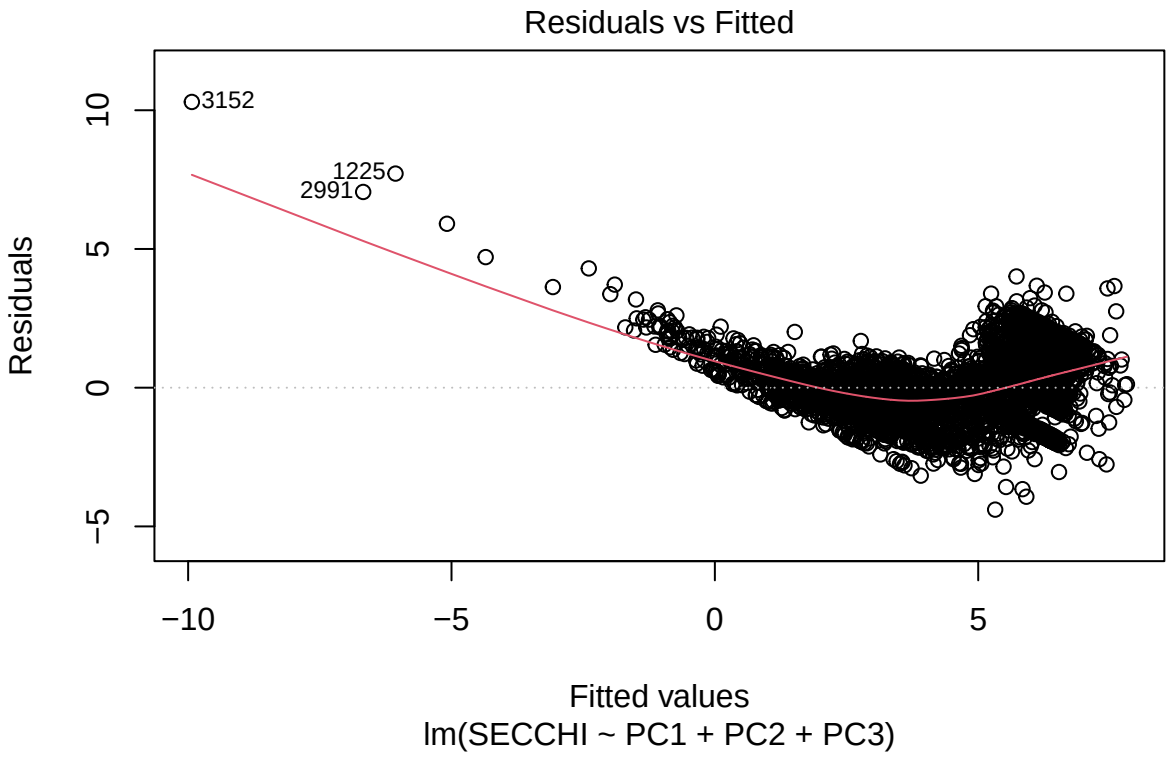
```
##                  PC1           PC2           PC3           PC4           PC5
## SECCHI  4.487283e-03  1.100908e-02 -2.522531e-01  6.604748e-01  7.969426e-05
## CHLA   -3.965672e-03 -1.330778e-01  6.480023e-01  2.497347e-01 -3.299131e-04
## TP     -1.501503e-02 -6.942096e-01 -1.282767e-01 -3.731892e-02 -2.359539e-06
## TN     -7.069220e-01  1.556143e-02 -2.511756e-03  3.584153e-03  1.137341e-05
## PC1     7.071068e-01 -8.359616e-10  4.188549e-09  5.312180e-10  8.964343e-06
## PC2    -5.451210e-13 -7.071068e-01  3.352517e-08  4.036352e-09  6.589738e-05
## PC3     2.229263e-13 -2.736142e-09 -7.071069e-01 -3.263468e-08 -3.303808e-04
## PC4    -1.129565e-14  1.318237e-10 -1.305915e-08  7.071068e-01  4.189780e-05
## PC5    -1.388629e-15  1.773370e-11 -1.089346e-09 -3.452035e-10  9.999999e-01
##                  PC6           PC7           PC8           PC9
## SECCHI -1.786442e-02  0.1005339534 -0.1140038252 -6.903455e-01
## CHLA   -3.086311e-03  0.2449775419  0.6592651827 -7.311564e-02
## TP      2.737045e-02  0.6552897074 -0.2286789763  1.324847e-01
## TN      7.063442e-01 -0.0217790611  0.0088584731 -2.291297e-02
## PC1     7.068369e-01 -0.0071227277  0.0084211121 -1.612288e-02
## PC2    -1.102387e-02 -0.6883565522  0.0988540311 -1.275602e-01
## PC3    -3.929722e-03  0.0698371231  0.6862826056  1.553168e-01
## PC4     1.564056e-02 -0.1457301537 -0.1384664091  6.777501e-01
## PC5    -1.512698e-05  0.0001492066  0.0004518917  6.293584e-05
```

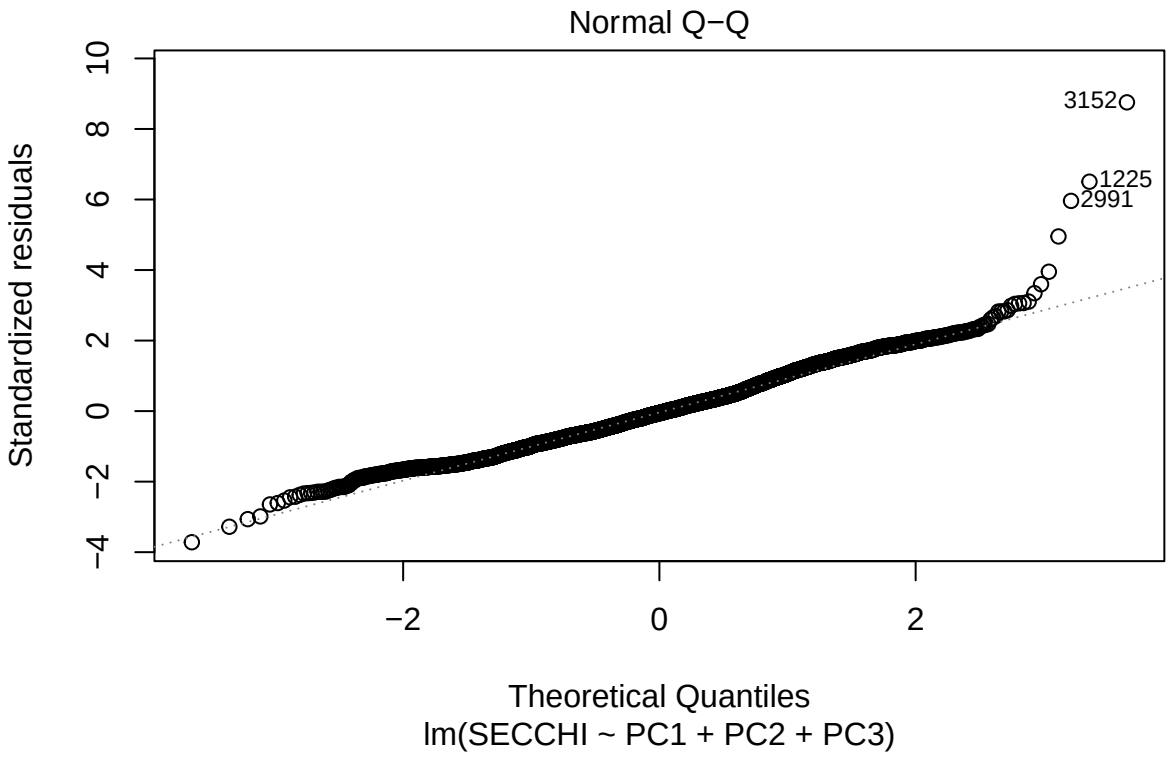
```
three.variables.pca.x <- three.variables.pca$x
df1$PC1 <- three.variables.pca.x[,1]
```

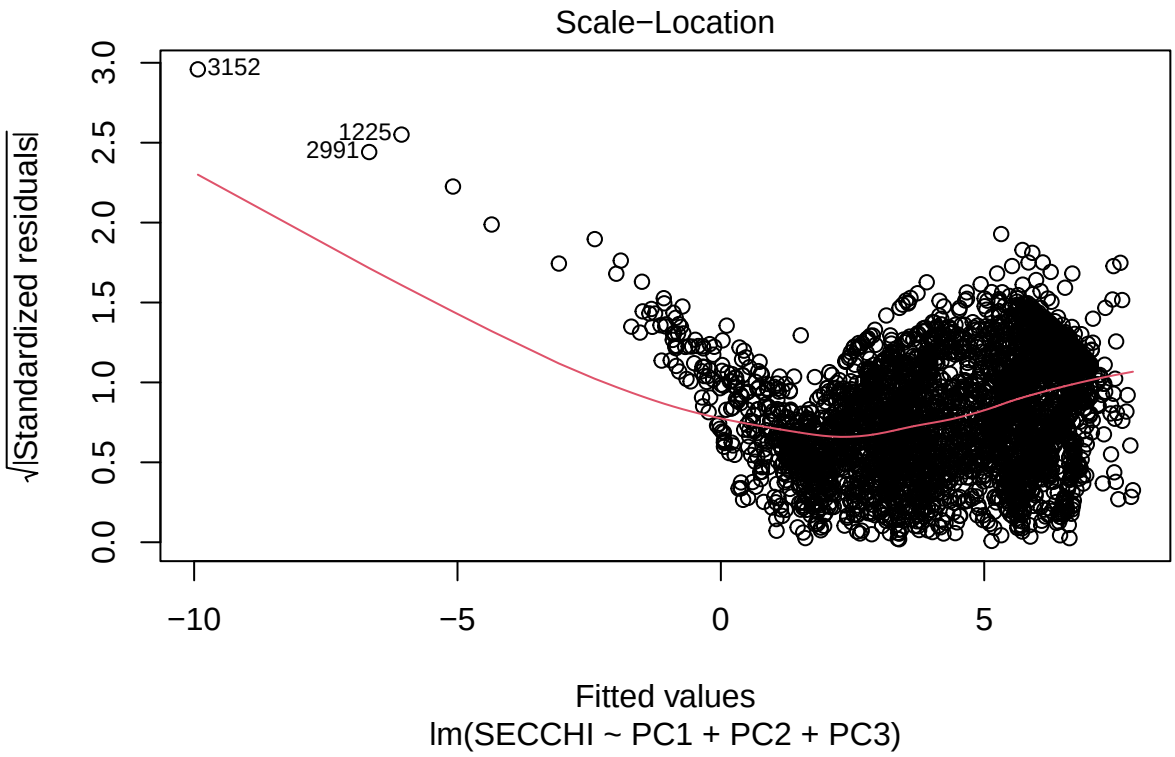
```r
df1$PC2 <- three.variables.pca.x[,2]
df1$PC3 <- three.variables.pca.x[,3]


full.lm.pca <- lm(data=df1, SECCHI~PC1+PC2+PC3)
summary(full.lm.pca)
```

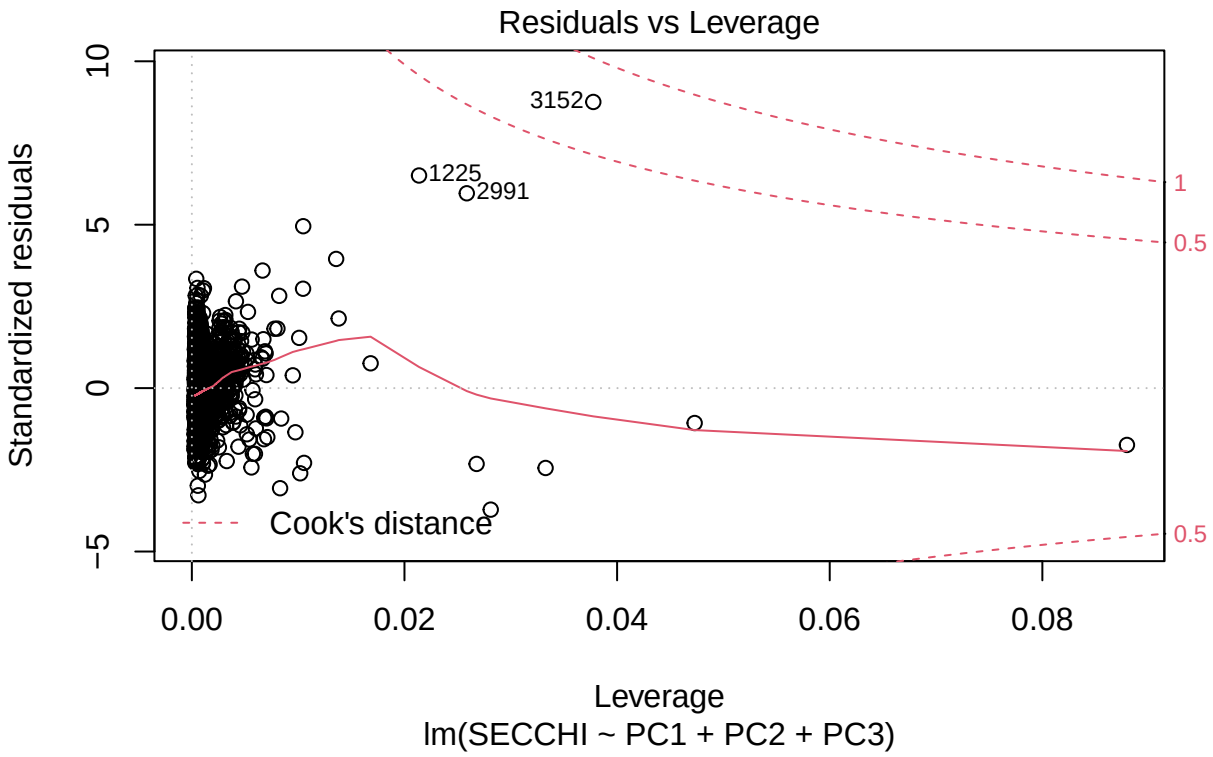
```
##
## Call:
## lm(formula = SECCHI ~ PC1 + PC2 + PC3, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3955 -0.8270 -0.0654  0.7360 10.2985
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.451e+00  1.945e-02 228.873  < 2e-16 ***
## PC1          4.487e-03  4.946e-05  90.726  < 2e-16 ***
## PC2          1.101e-02  1.937e-03   5.684 1.41e-08 ***
## PC3         -2.523e-01  6.779e-03 -37.208  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 3798 degrees of freedom
## Multiple R-squared:  0.7175, Adjusted R-squared:  0.7173
## F-statistic:  3216 on 3 and 3798 DF,  p-value: < 2.2e-16
```

```r
plot(full.lm.pca)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SECCHI ~ PC1 + PC2 + PC3)
3152
1225
2991
10
5
0
−5
−10
−5
0
5


Normal Q−Q
Standardized residuals
10
8
6
4
2
0
−2
−4
3152
1225
2991
−2
0
2
Theoretical Quantiles
lm(SECCHI ~ PC1 + PC2 + PC3)


Scale−Location
√|Standardized residuals|
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3152
1225
2991
−10
−5
0
5
Fitted values
lm(SECCHI ~ PC1 + PC2 + PC3)


Residuals vs Leverage
Standardized residuals
10
5
0
−5
3152
1225
2991
1
0.5
0.5
Cook's distance
0.00
0.02
0.04
0.06
0.08
Leverage
lm(SECCHI ~ PC1 + PC2 + PC3)