MSC_609_Code_Grp_4

Joshua Man Yin Wong

2024-11-20

Generation of Histogram Plot

```
## Warning: package 'patchwork' was built under R version 4.4.2
```

```
##
## Attaching package: 'patchwork'
```

```
## The following object is masked from 'package:cowplot':
##
##     align_plots
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 25 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 23 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 57 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

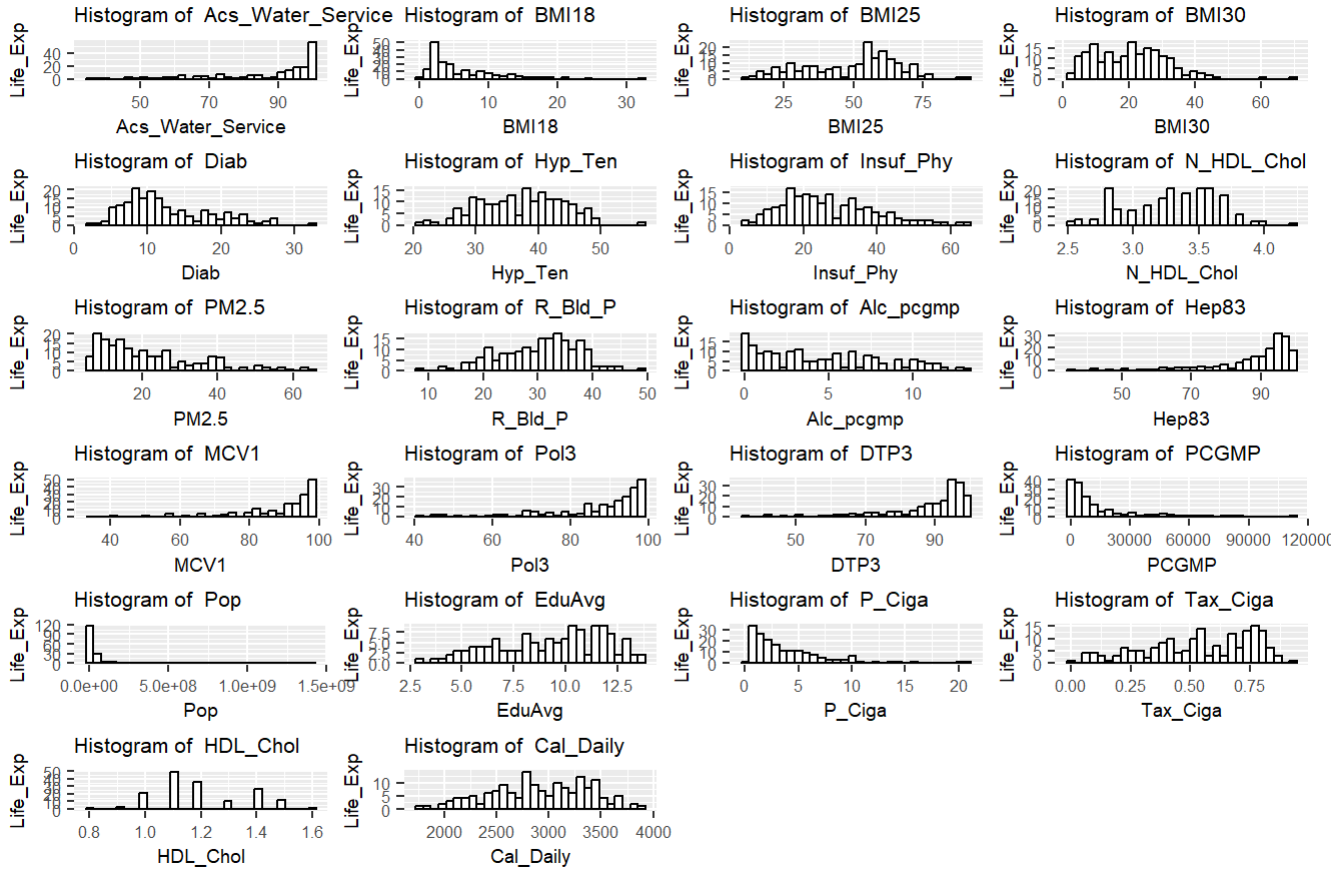
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 21 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

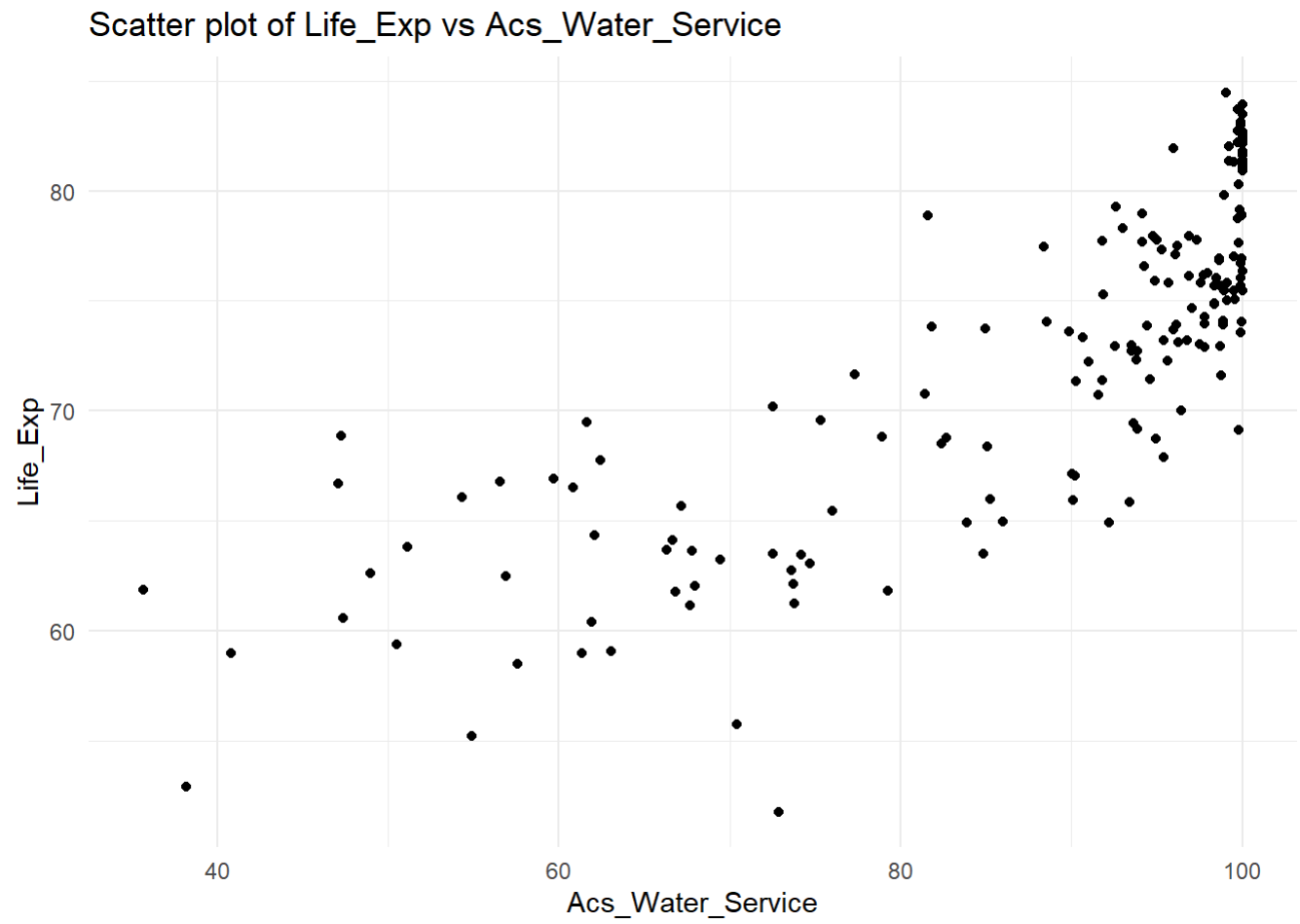
```
## Warning: Removed 25 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

Histogram of all Predictors

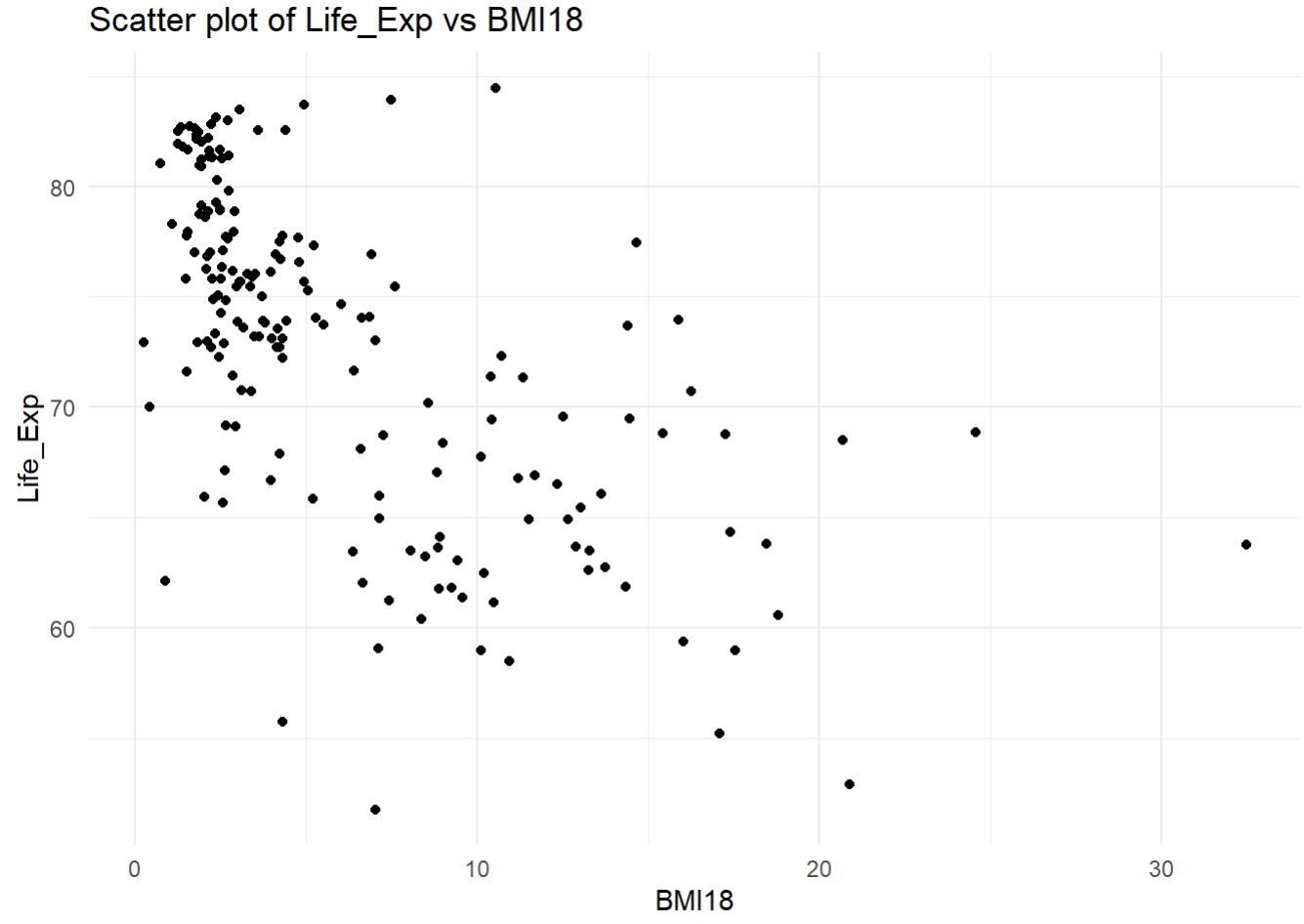


Initial Data Visualization

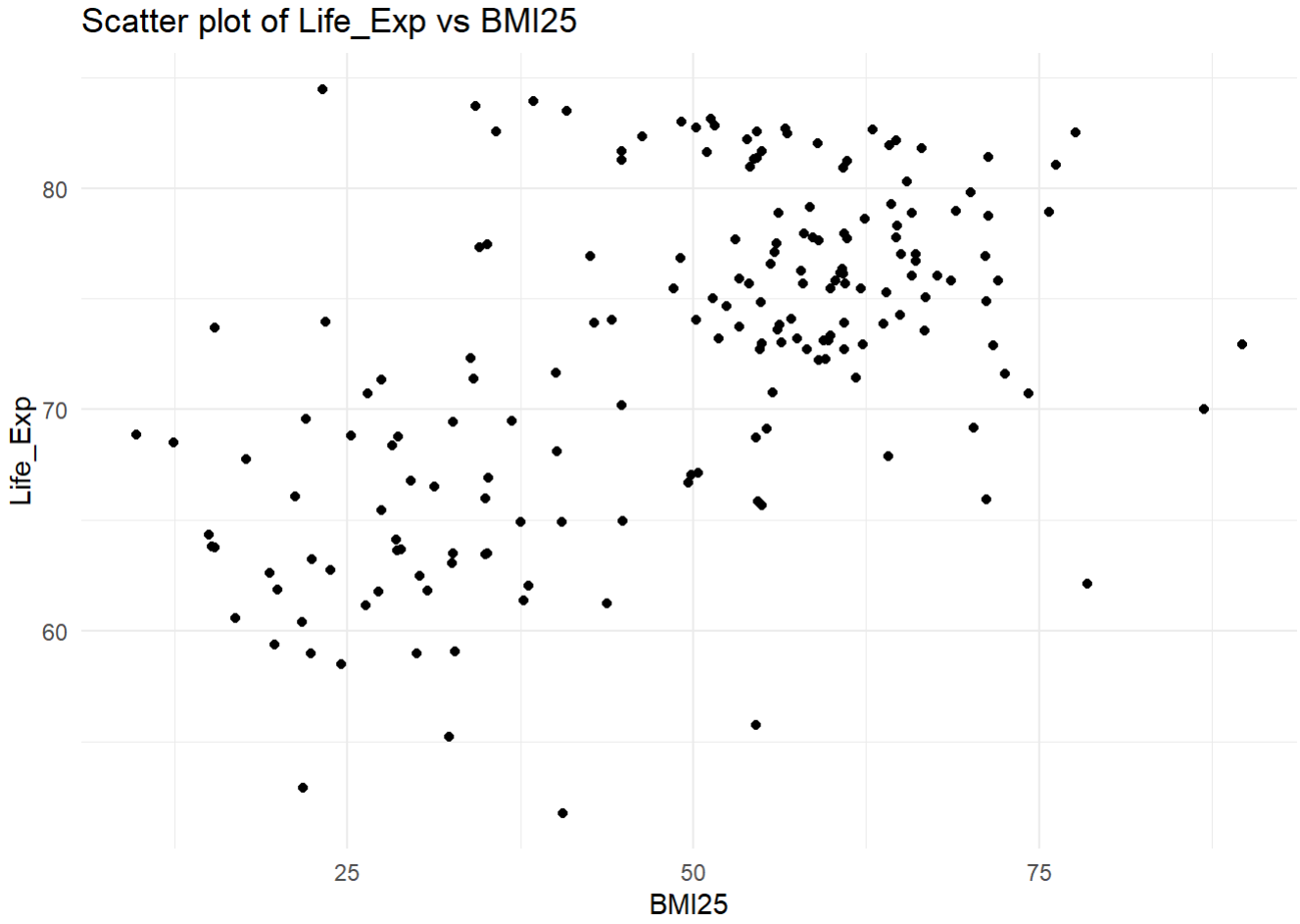
```
## [[1]]
## NULL
##
## [[2]]
```


Scatter plot of Life_Exp vs Acs_Water_Service
Life_Exp
80
70
60
40
60
80
100
Acs_Water_Service

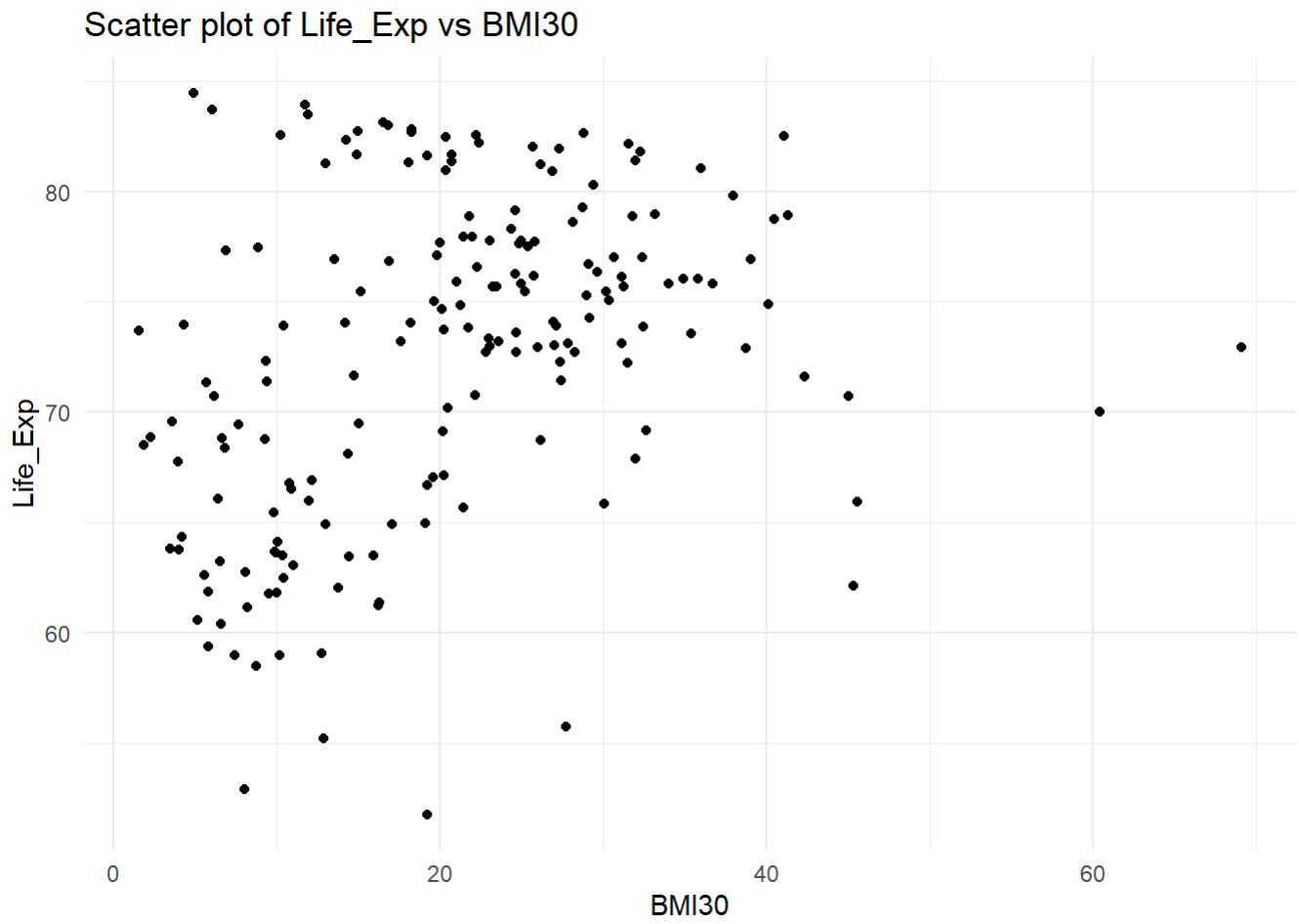
```
##
## [[3]]
```


Scatter plot of Life_Exp vs BMI18
Life_Exp
80
70
60
0
10
20
30
BMI18

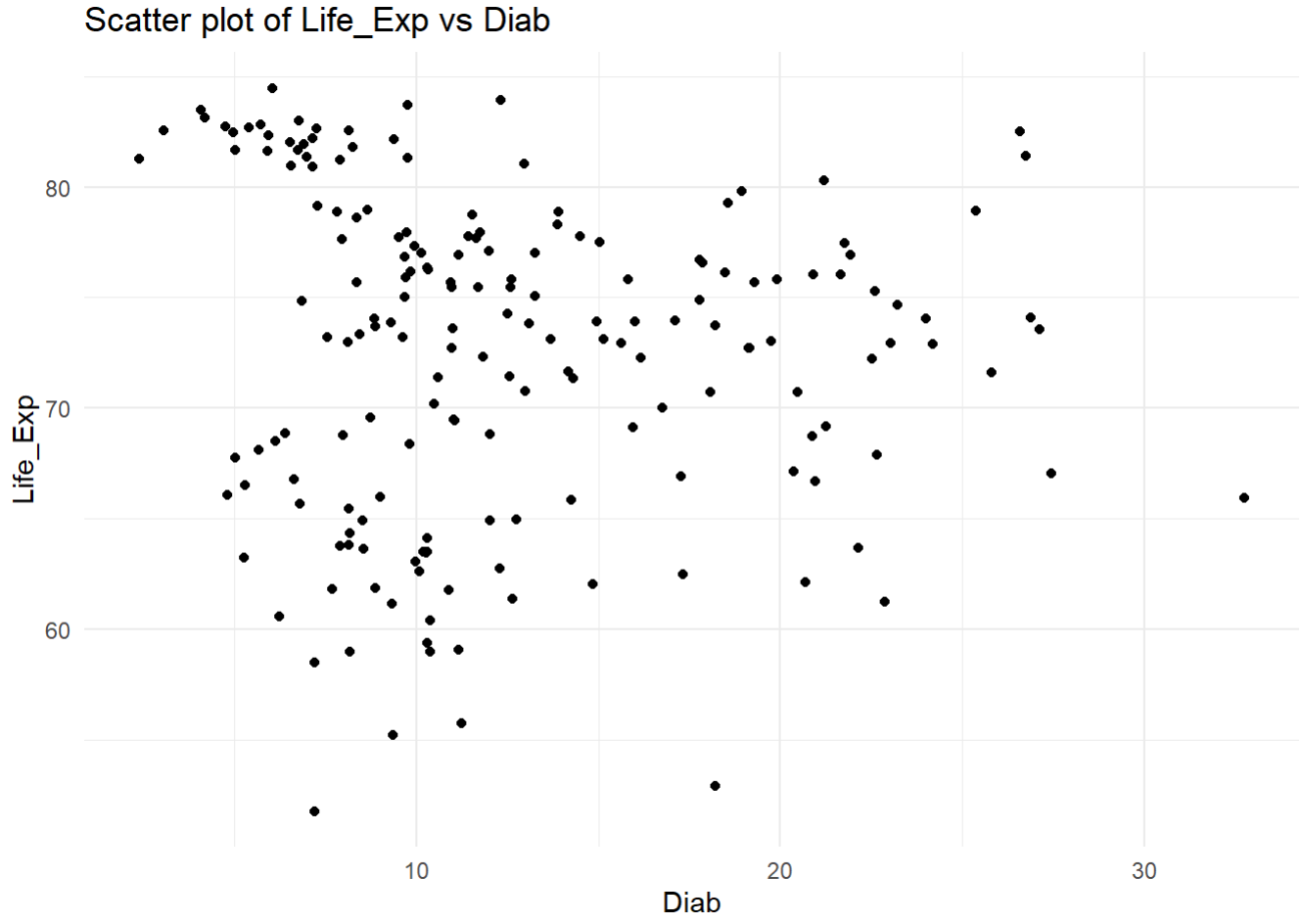
```
##
## [[4]]
```


Scatter plot of Life_Exp vs BMI25
Life_Exp
80
70
60
25
50
75
BMI25

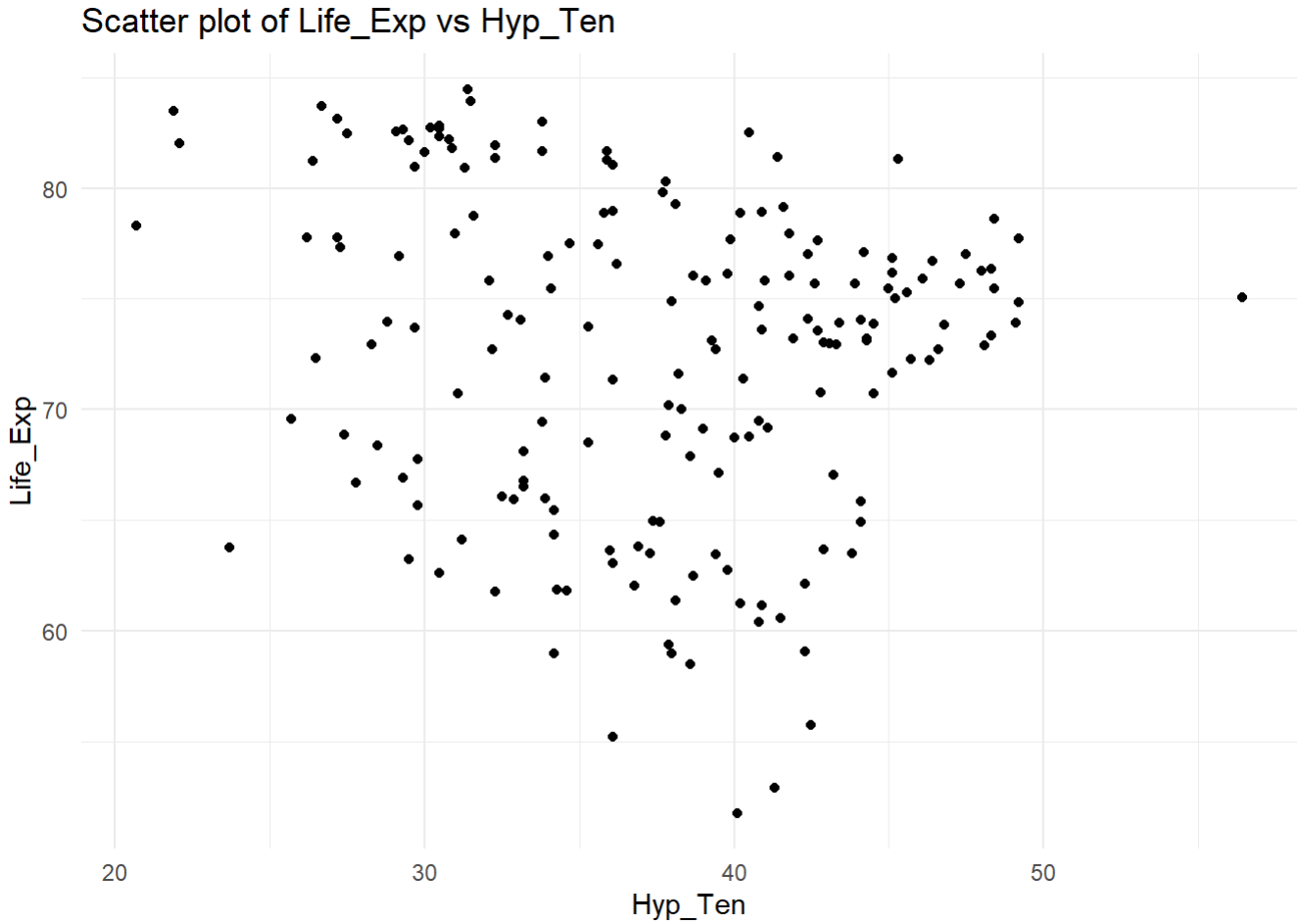
```
##
## [[5]]
```


Scatter plot of Life_Exp vs BMI30
Life_Exp
80
70
60
0
20
40
60
BMI30

```
##
## [[6]]
```


Scatter plot of Life_Exp vs Diab
Life_Exp
80
70
60
10
20
30
Diab

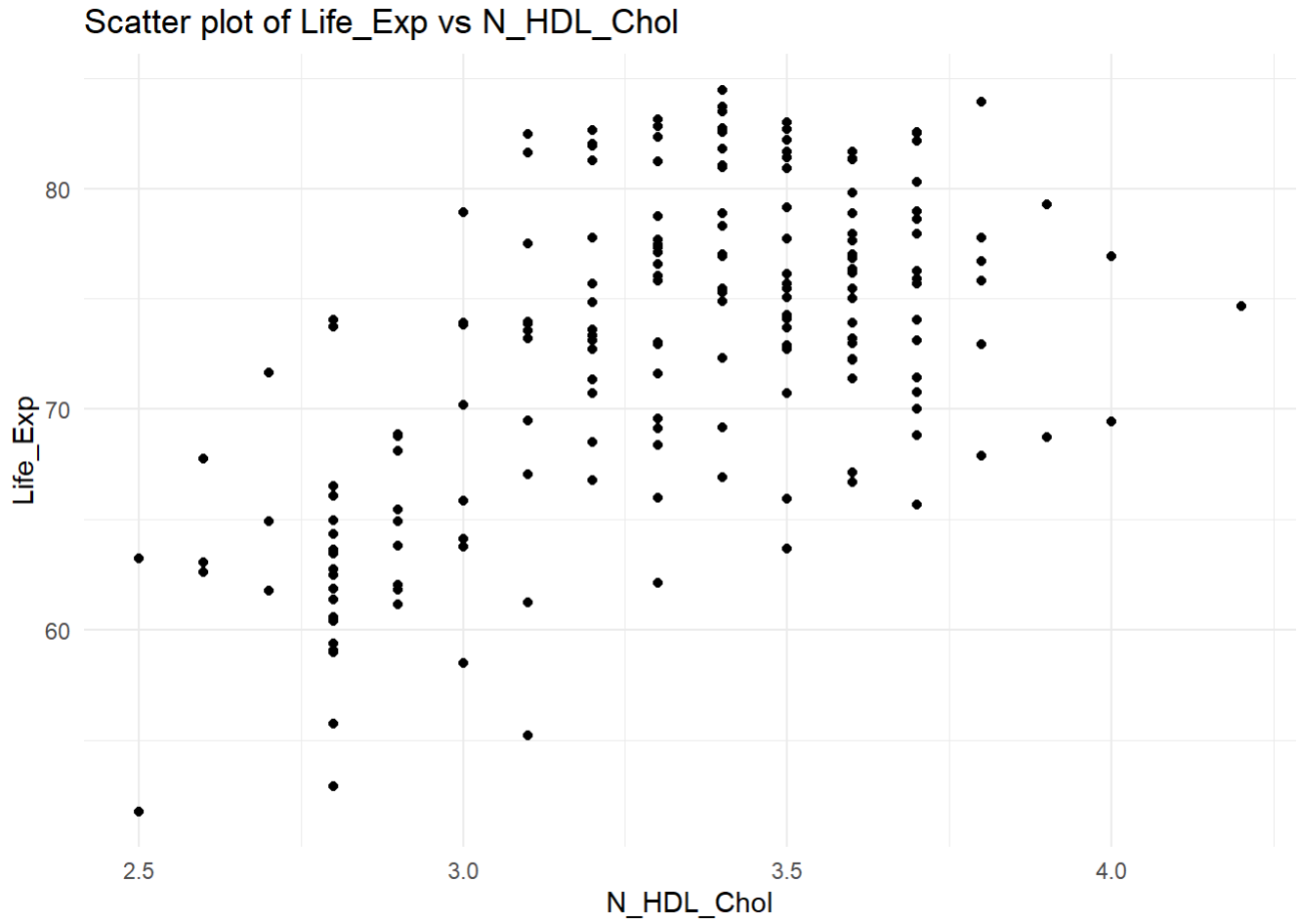
```
##
## [[7]]
```


Scatter plot of Life_Exp vs Hyp_Ten
Life_Exp
80
70
60
20
30
40
50
Hyp_Ten

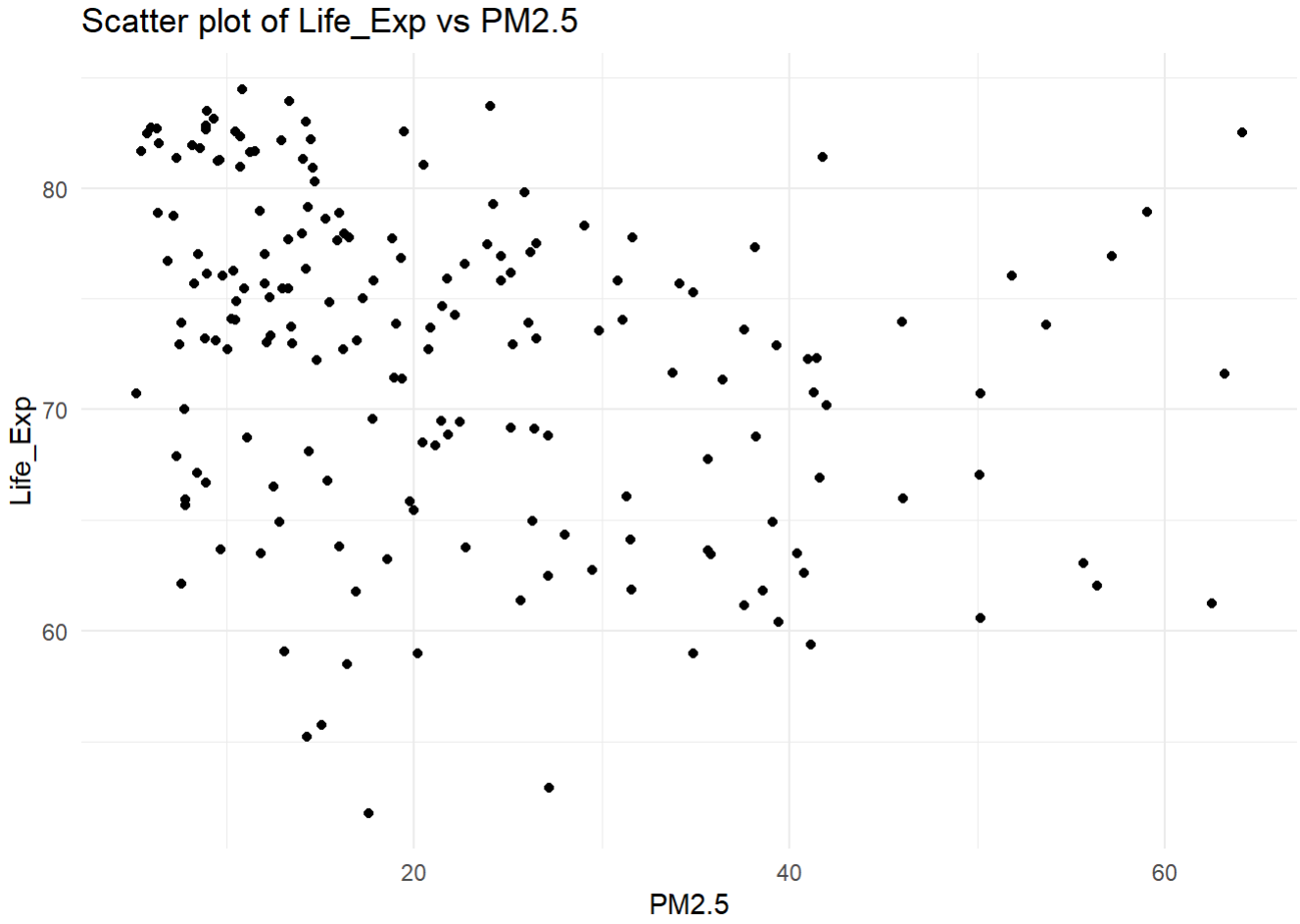
```
##
## [[8]]
```


Scatter plot of Life_Exp vs Insuf_Phy
Life_Exp
80
70
60
20
40
60
Insuf_Phy

```
##
## [[9]]
```


Scatter plot of Life_Exp vs N_HDL_Chol
Life_Exp
80
70
60
2.5
3.0
3.5
4.0
N_HDL_Chol

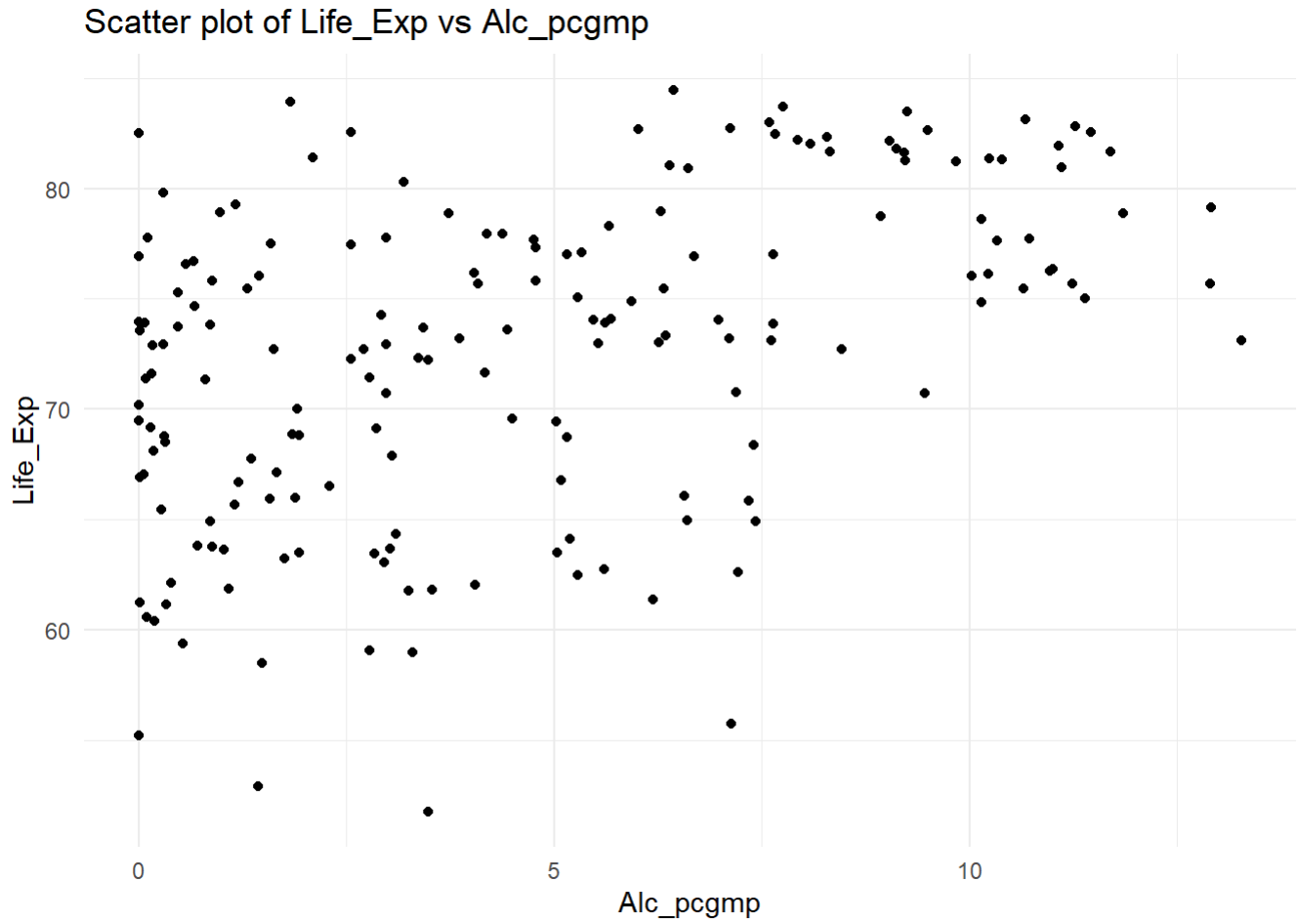
```
##
## [[10]]
```



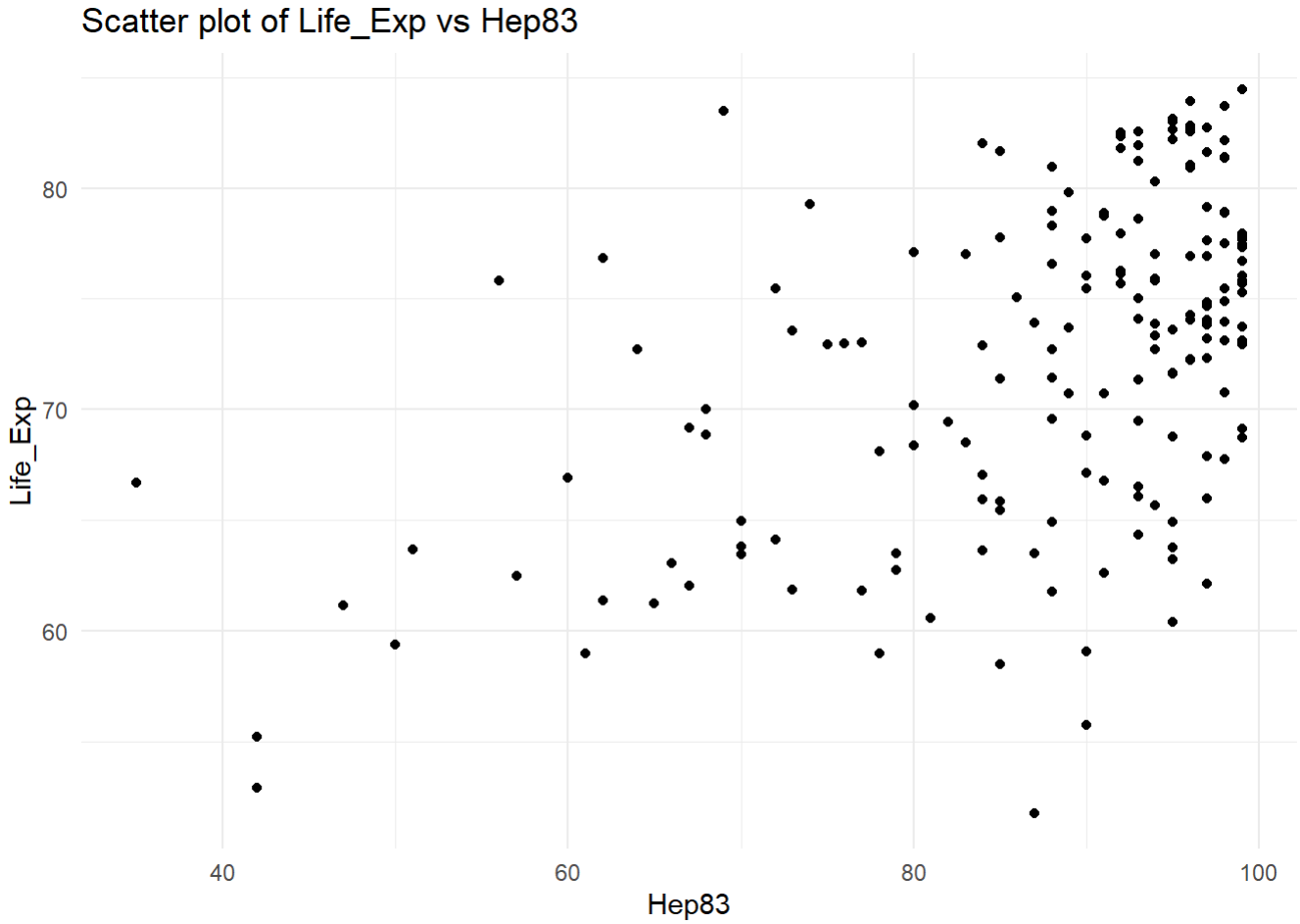

```
##
## [[11]]
```


Scatter plot of Life_Exp vs R_Bld_P
Life_Exp
80
70
60
10
20
30
40
50
R_Bld_P

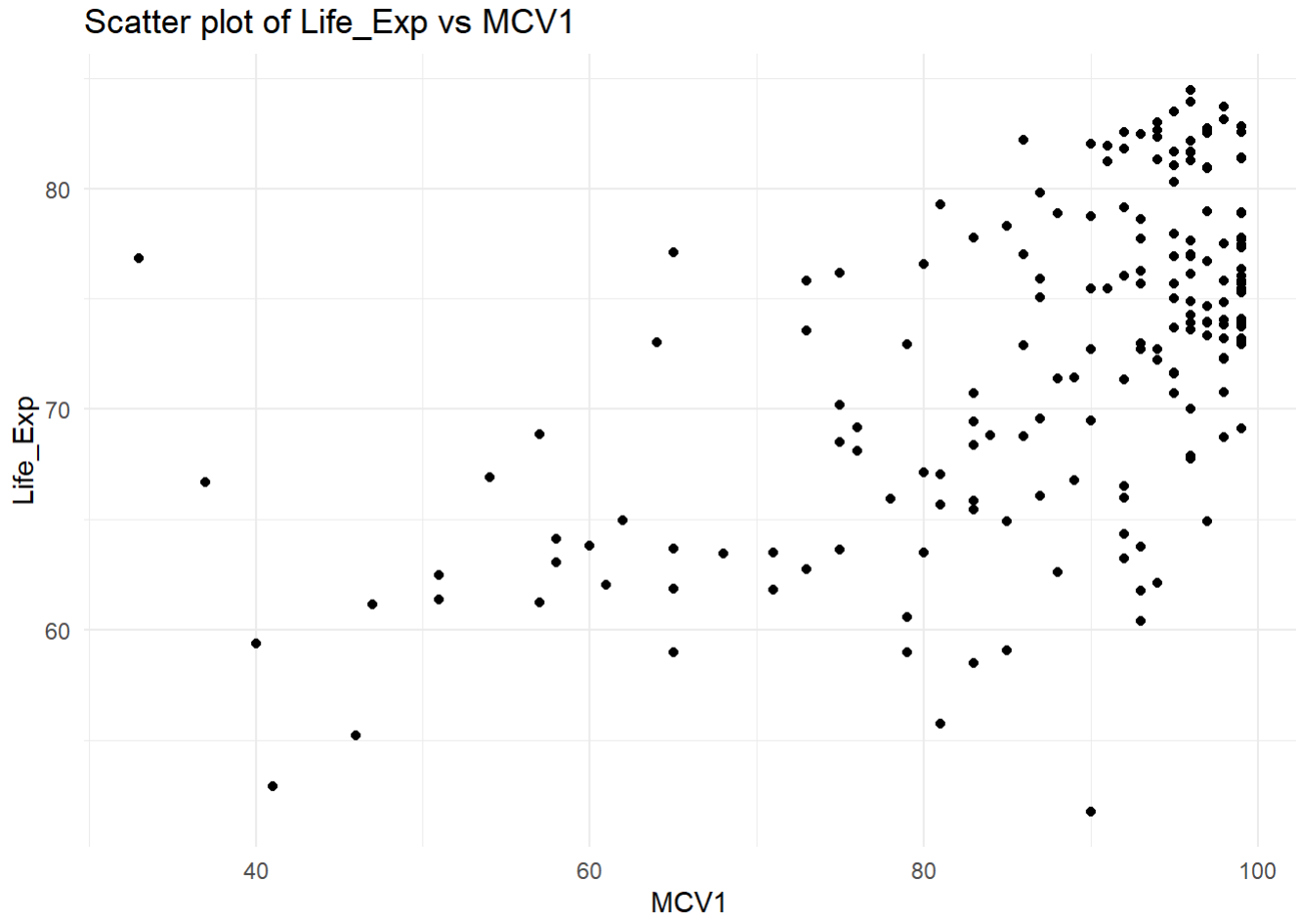
```
##
## [[12]]
```


Scatter plot of Life_Exp vs Alc_pcgmp
Life_Exp
80
70
60
0
5
10
Alc_pcgmp

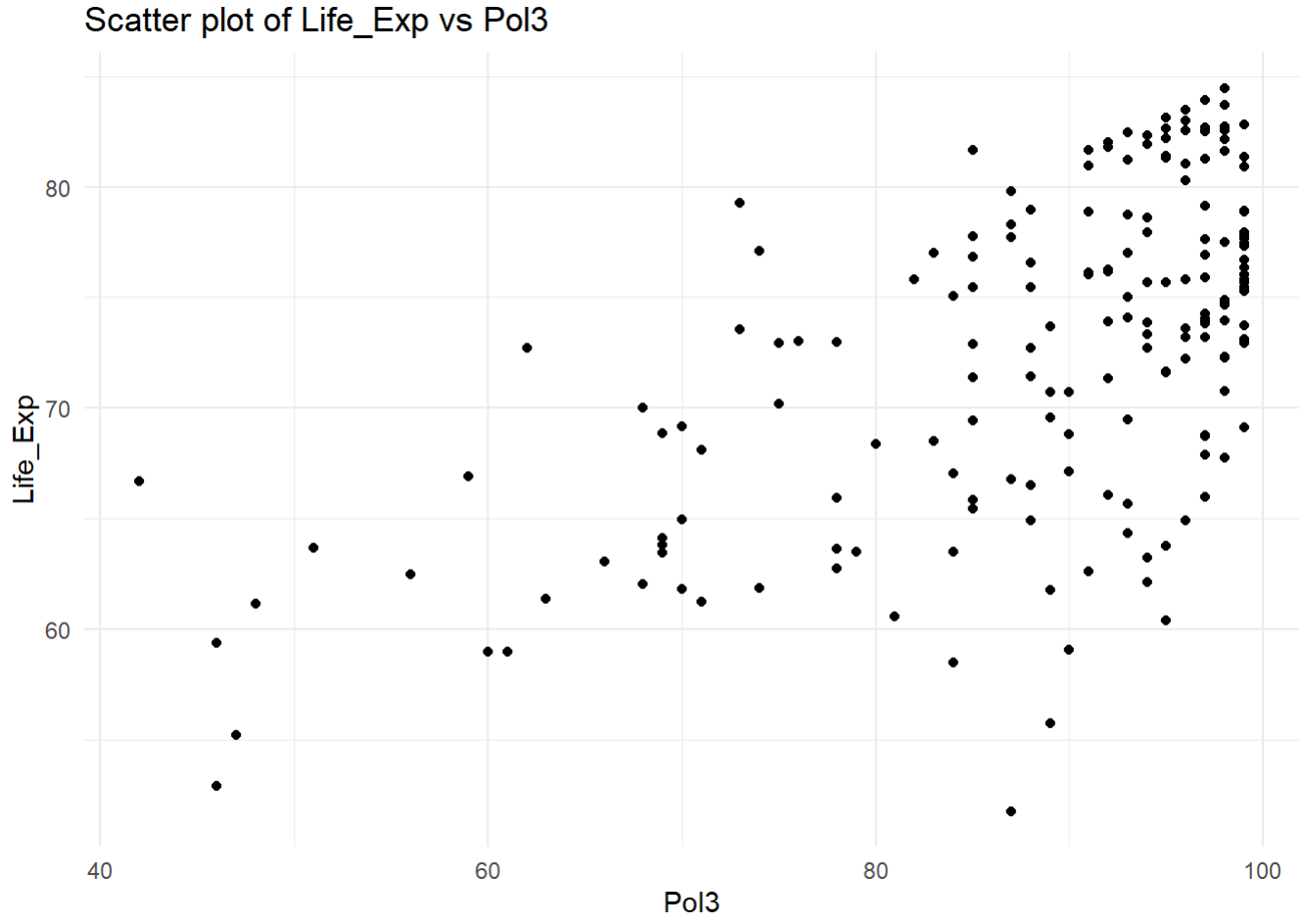
```
##
## [[13]]
```


Scatter plot of Life_Exp vs Hep83
Life_Exp
80
70
60
40
60
80
100
Hep83

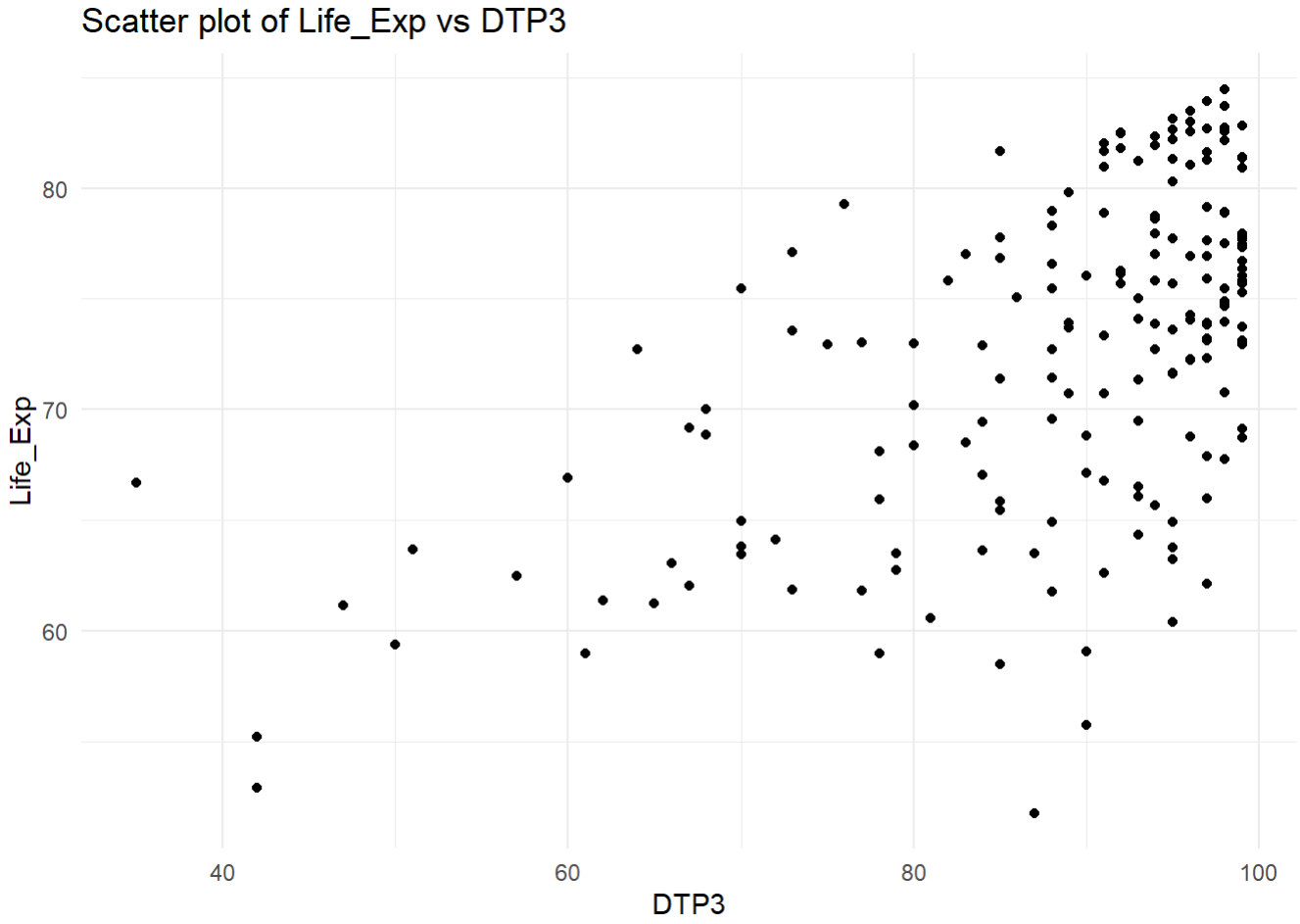
```
##
## [[14]]
```


Scatter plot of Life_Exp vs MCV1
Life_Exp
80
70
60
40
60
80
100
MCV1

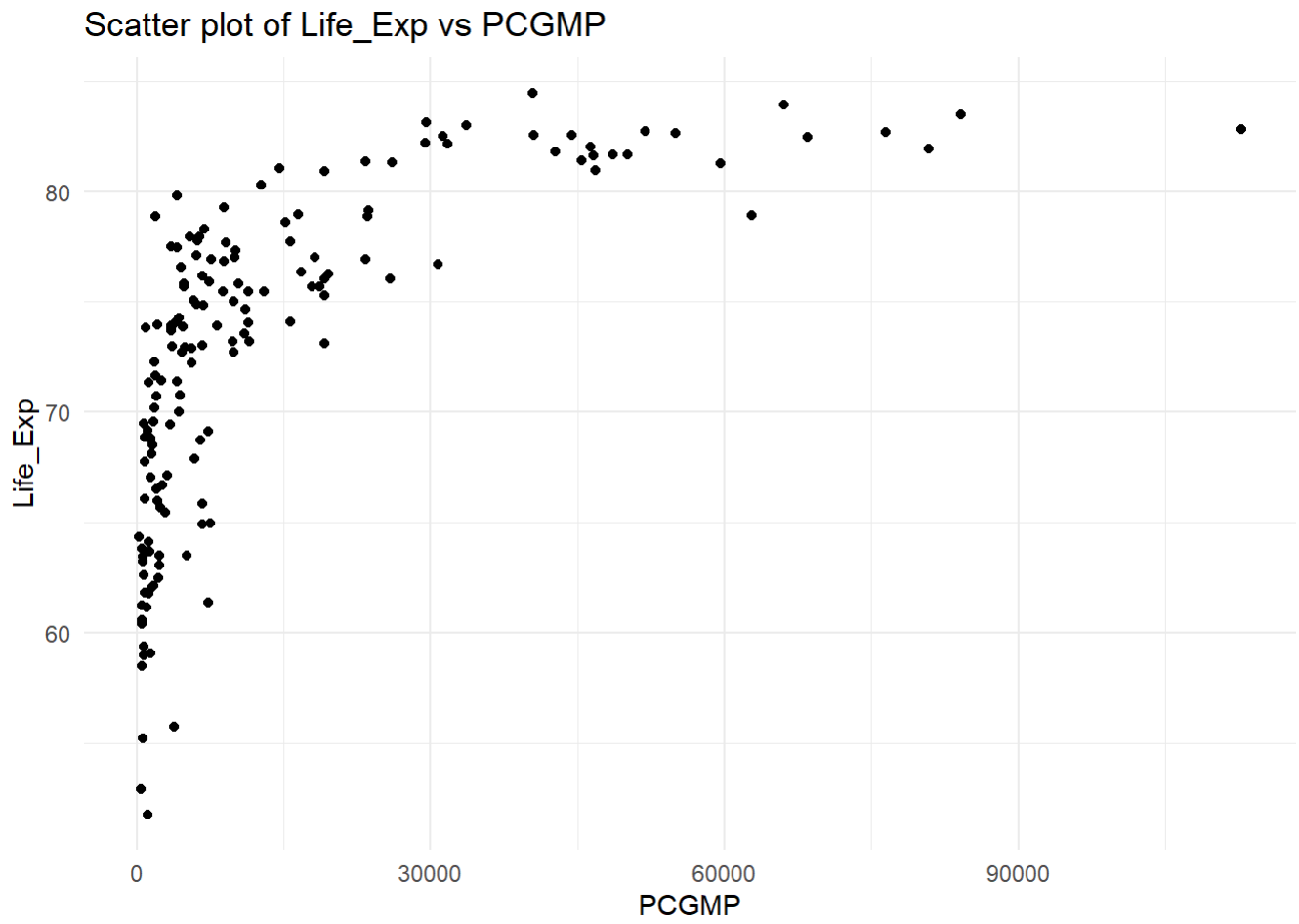
```
##
## [[15]]
```


Scatter plot of Life_Exp vs Pol3
Life_Exp
80
70
60
40
60
80
100
Pol3

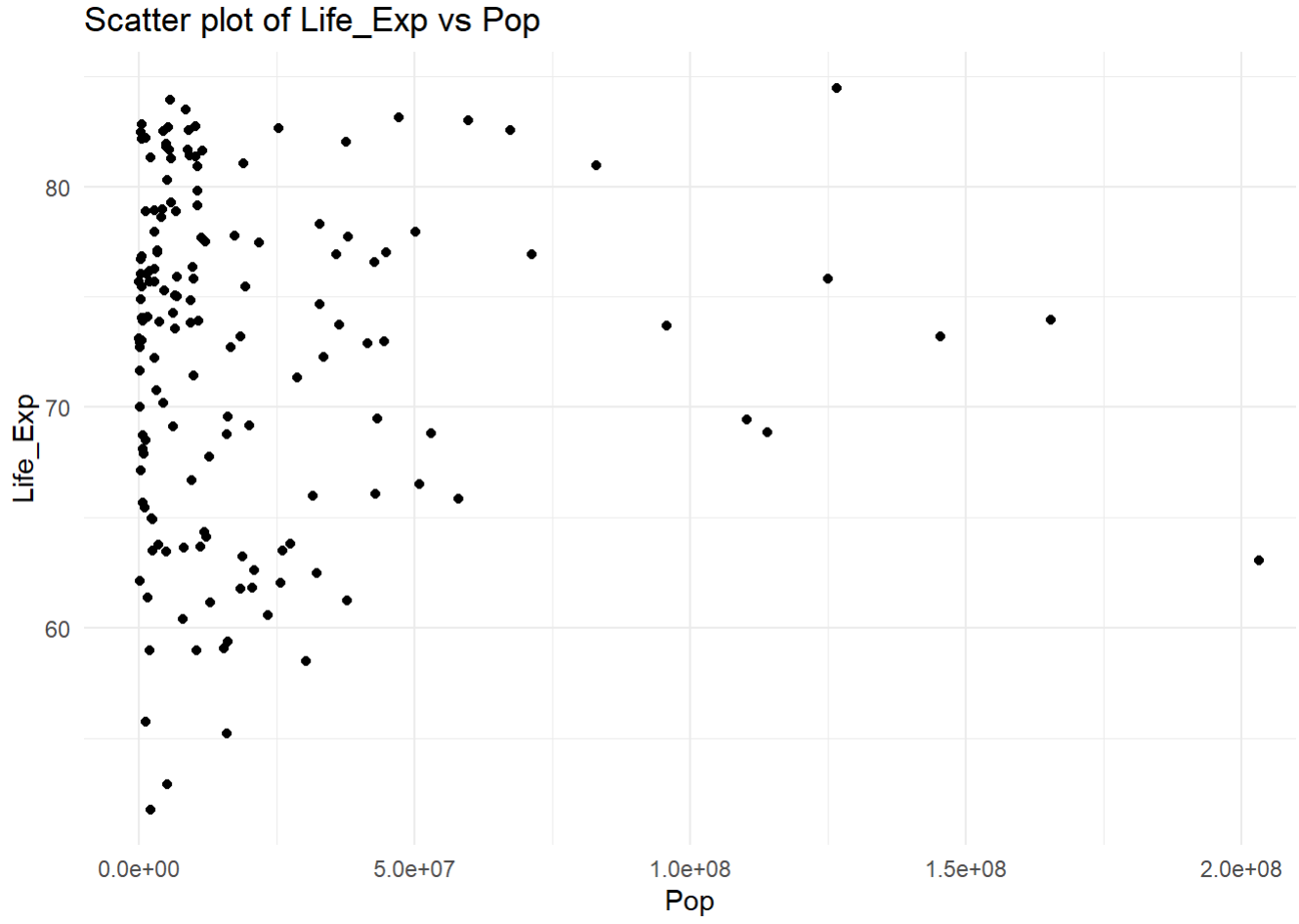
```
##
## [[16]]
```


Scatter plot of Life_Exp vs DTP3
Life_Exp
80
70
60
40
60
80
100
DTP3

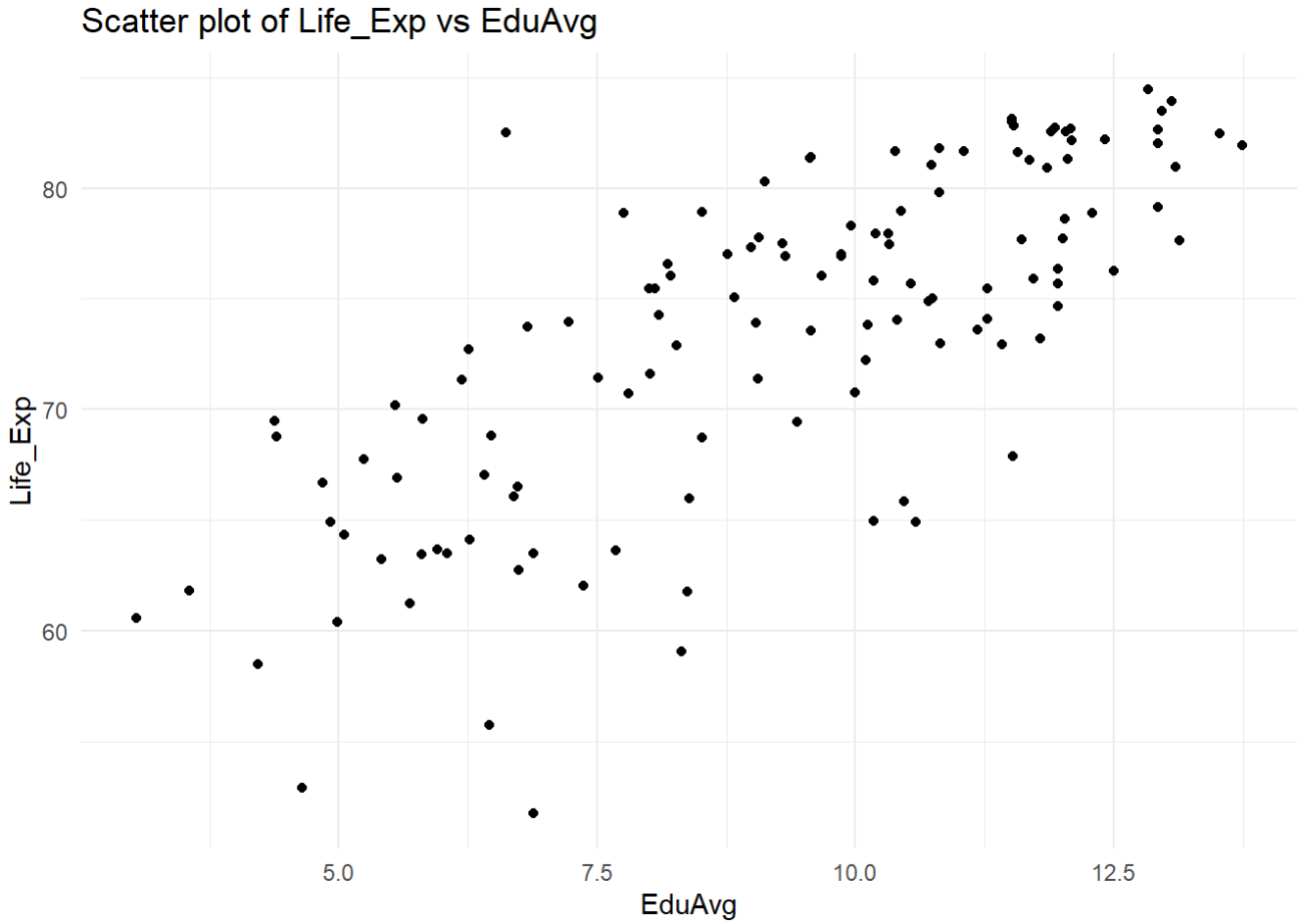
```
##
## [[17]]
```


Scatter plot of Life_Exp vs PCGMP
Life_Exp
80
70
60
0
30000
60000
90000
PCGMP

```
##
## [[18]]
```


Scatter plot of Life_Exp vs Pop
Life_Exp
80
70
60
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Pop

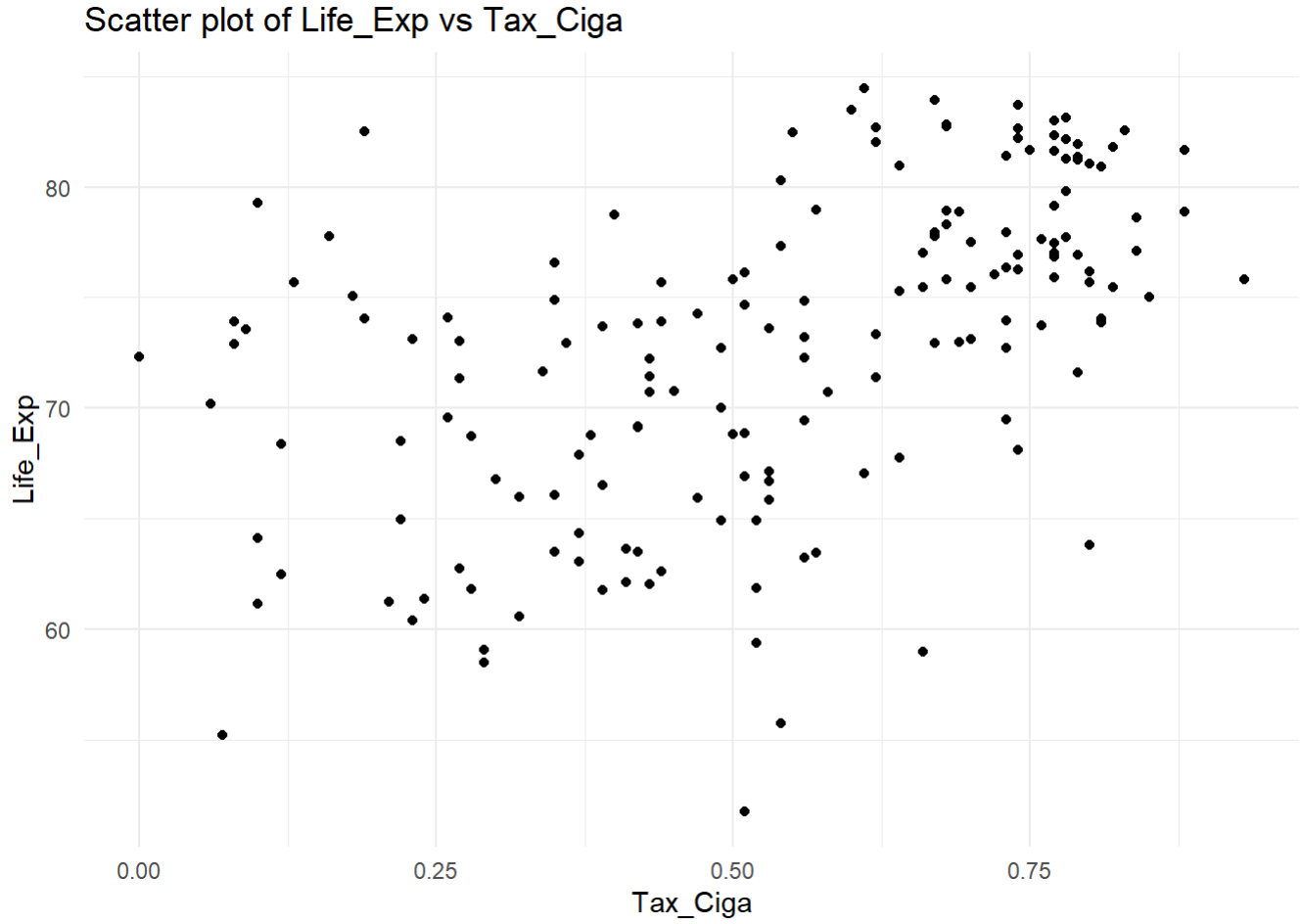
```
##
## [[19]]
```


Scatter plot of Life_Exp vs EduAvg
Life_Exp
80
70
60
5.0
7.5
10.0
12.5
EduAvg

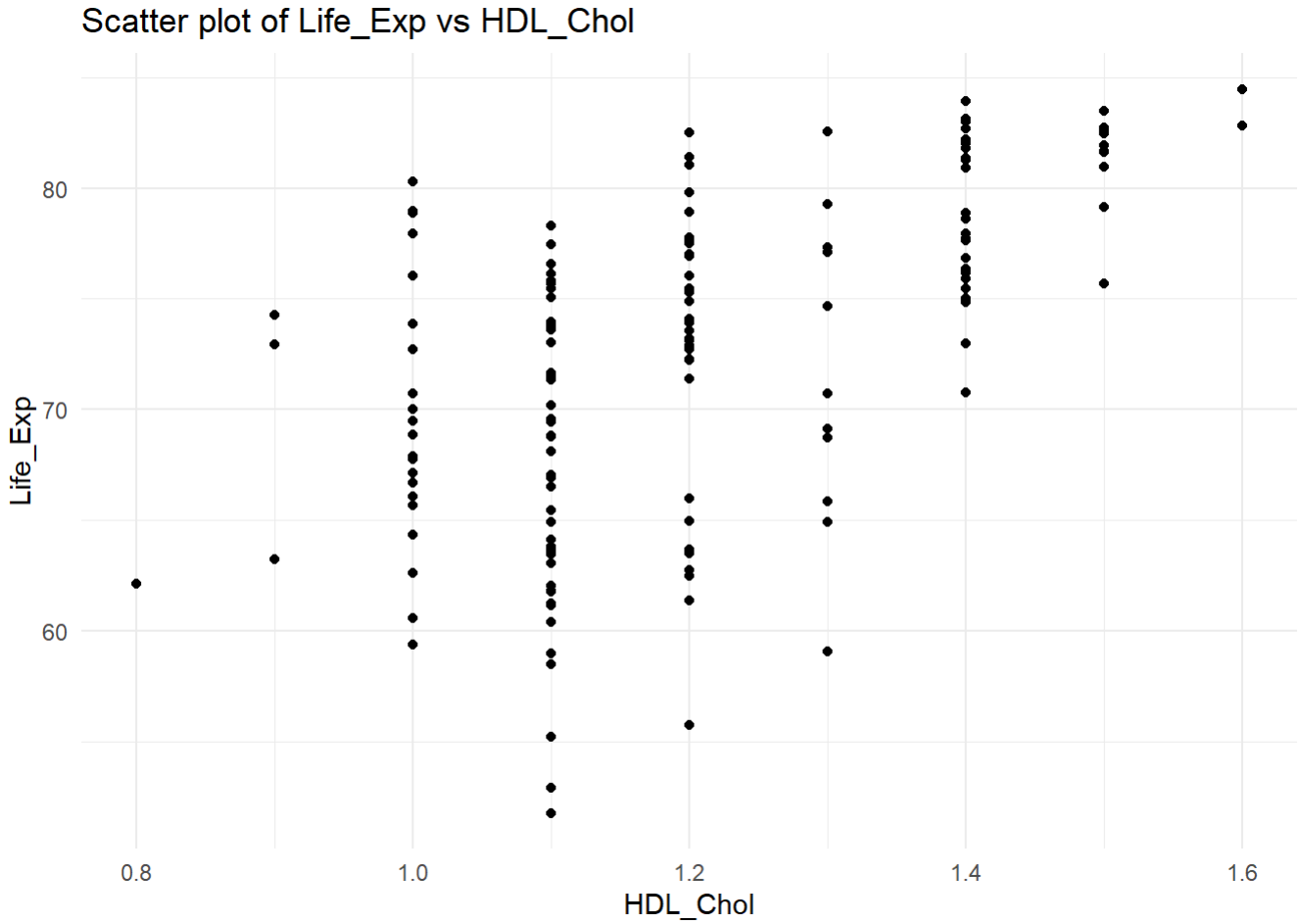
```
##
## [[20]]
```


Scatter plot of Life_Exp vs P_Ciga
Life_Exp
80
70
60
0
5
10
15
20
P_Ciga

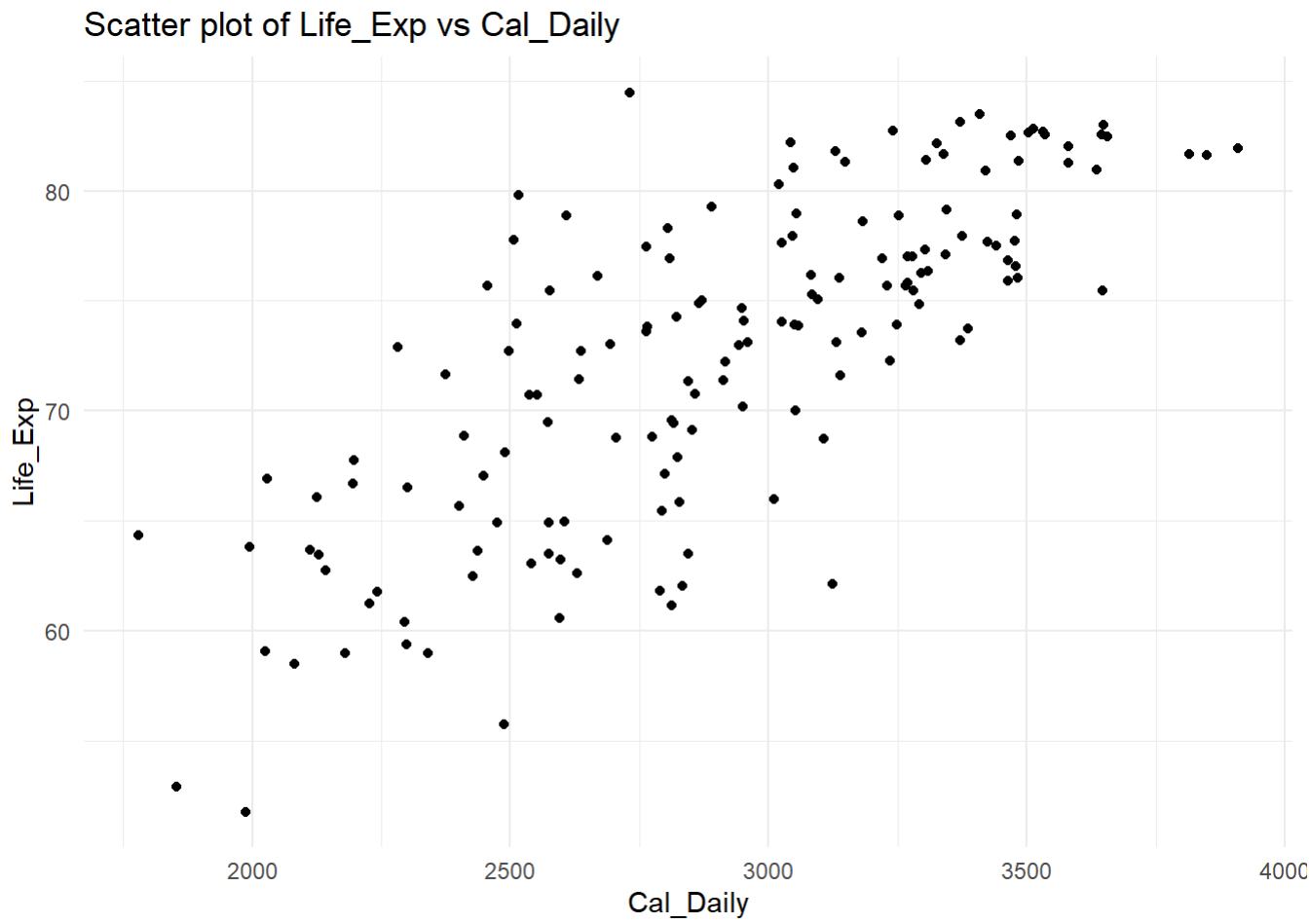
```
##
## [[21]]
```


Scatter plot of Life_Exp vs Tax_Ciga
Life_Exp
80
70
60
0.00
0.25
0.50
0.75
Tax_Ciga

```
##
## [[22]]
```


Scatter plot of Life_Exp vs HDL_Chol
Life_Exp
80
70
60
0.8
1.0
1.2
1.4
1.6
HDL_Chol

```
##
## [[23]]
```



Approach 1: Keep only complete cases for analysis

184

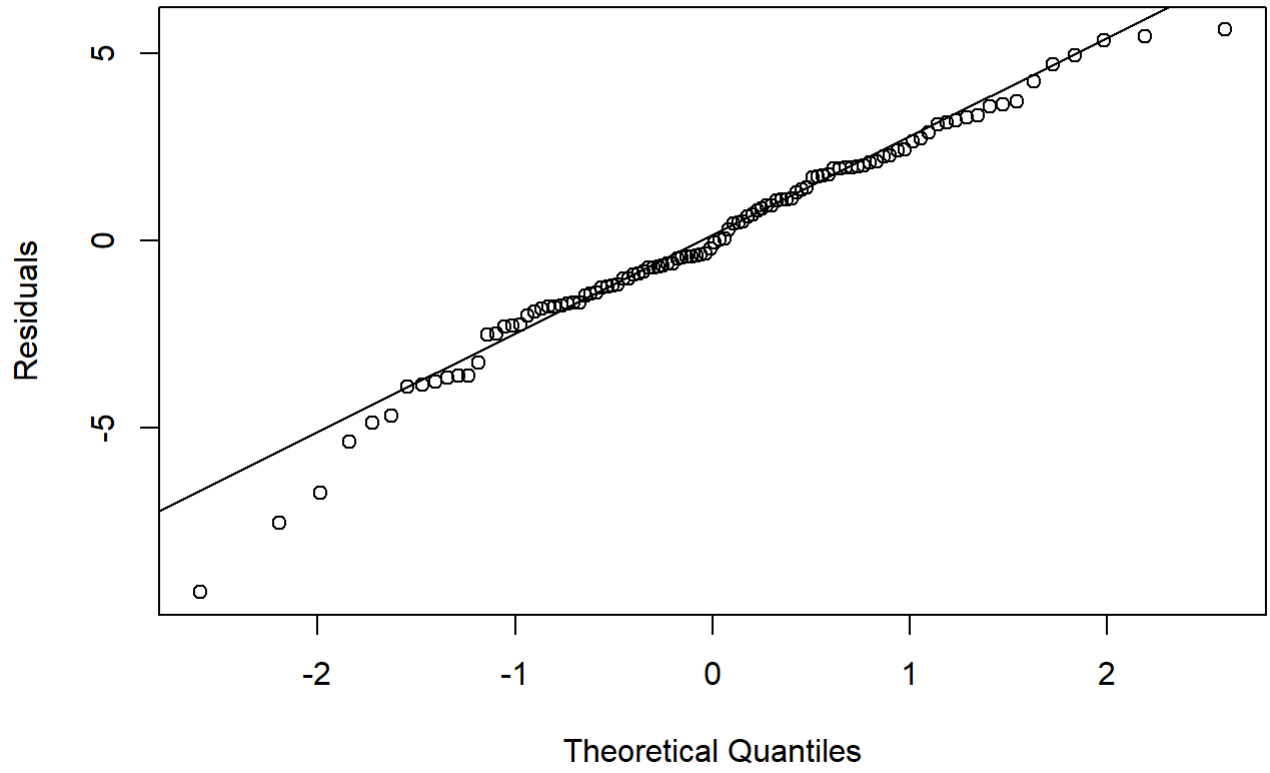
```
# summary(df_AdjP)

#  -- initial data = df_184
com184_lmod <- lm(Life_Exp ~. ,data = df_184)

test_mod <- com184_lmod

summary(test_mod)
```

```
##
## Call:
## lm(formula = Life_Exp ~ ., data = df_184)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3947 -1.6150 -0.1291  1.9388  5.6356
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.122e+01  6.188e+00   5.045 2.62e-06 ***
## Acs_Water_Service  6.309e-02  4.912e-02   1.284  0.20255
## BMI18              2.389e-01  1.815e-01   1.316  0.19174
## BMI25              2.774e-01  1.202e-01   2.308  0.02349 *
## BMI30             -2.453e-01  1.478e-01  -1.660  0.10072
## Diab              -1.950e-01  8.702e-02  -2.241  0.02770 *
## Hyp_Ten           -2.770e-01  1.548e-01  -1.789  0.07722 .
## Insuf_Phy          9.774e-02  4.156e-02   2.352  0.02104 *
## N_HDL_Chol         3.698e+00  1.280e+00   2.889  0.00493 **
## PM2.5              6.902e-03  3.118e-02   0.221  0.82535
## R_Bld_P            6.323e-02  1.386e-01   0.456  0.64946
## Alc_pcgmp         -2.670e-01  1.684e-01  -1.585  0.11678
## Hep83              6.708e-02  9.246e-02   0.726  0.47018
## MCV1               7.815e-02  8.395e-02   0.931  0.35458
## Pol3               3.545e-02  1.511e-01   0.235  0.81504
## DTP3              -1.188e-01  1.540e-01  -0.771  0.44274
## PCGMP              6.264e-05  3.014e-05   2.078  0.04079 *
## Pop               -4.928e-10  1.860e-09  -0.265  0.79173
## EduAvg             3.907e-01  2.669e-01   1.464  0.14691
## P_Ciga            -3.299e-02  1.417e-01  -0.233  0.81654
## Tax_Ciga           3.721e+00  2.072e+00   1.796  0.07619 .
## HDL_Chol           1.858e+00  3.954e+00   0.470  0.63965
## Cal_Daily          3.080e-03  1.155e-03   2.667  0.00920 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.08 on 83 degrees of freedom
##   (77 observations deleted due to missingness)
## Multiple R-squared:  0.8623, Adjusted R-squared:  0.8257
## F-statistic: 23.62 on 22 and 83 DF,  p-value: < 2.2e-16
```

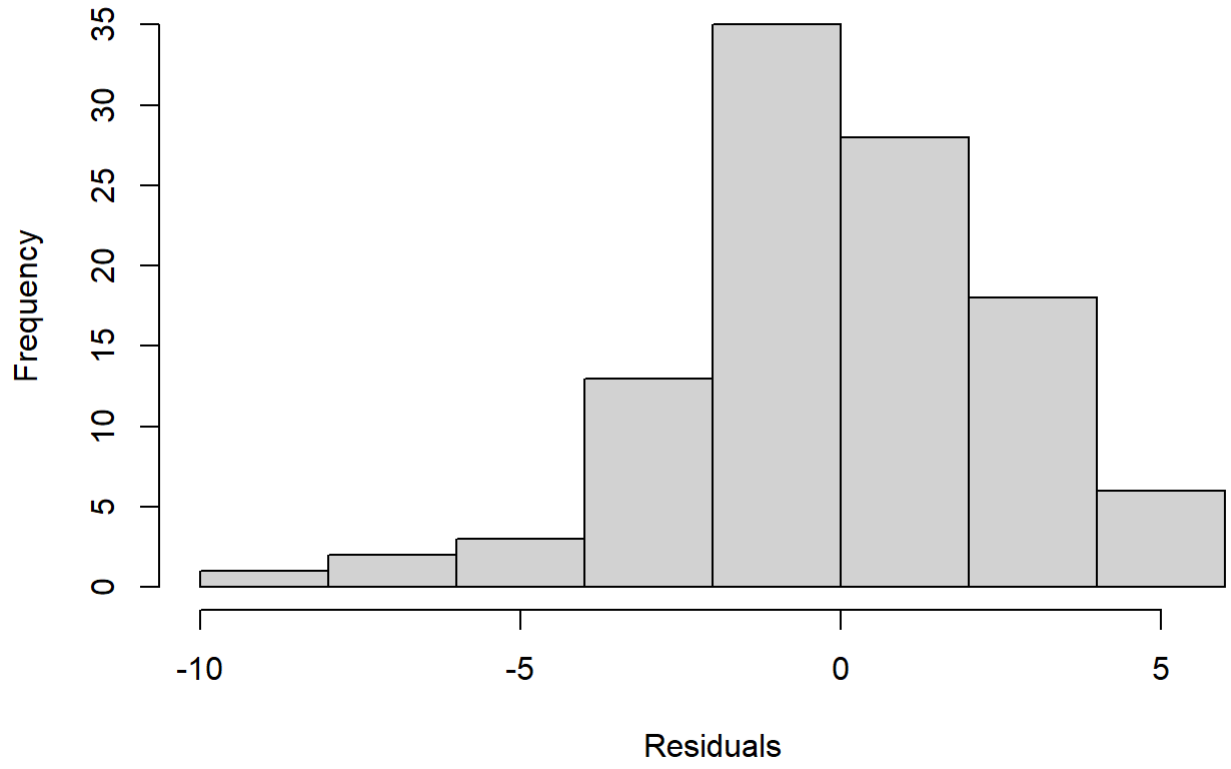
```
cat("AIC test: ", AIC(test_mod, k = 2))
```

```
## AIC test:  561.3457
```

```
qqnorm(residuals(test_mod), ylab = "Residuals", main="")
qqline(residuals(test_mod))
```



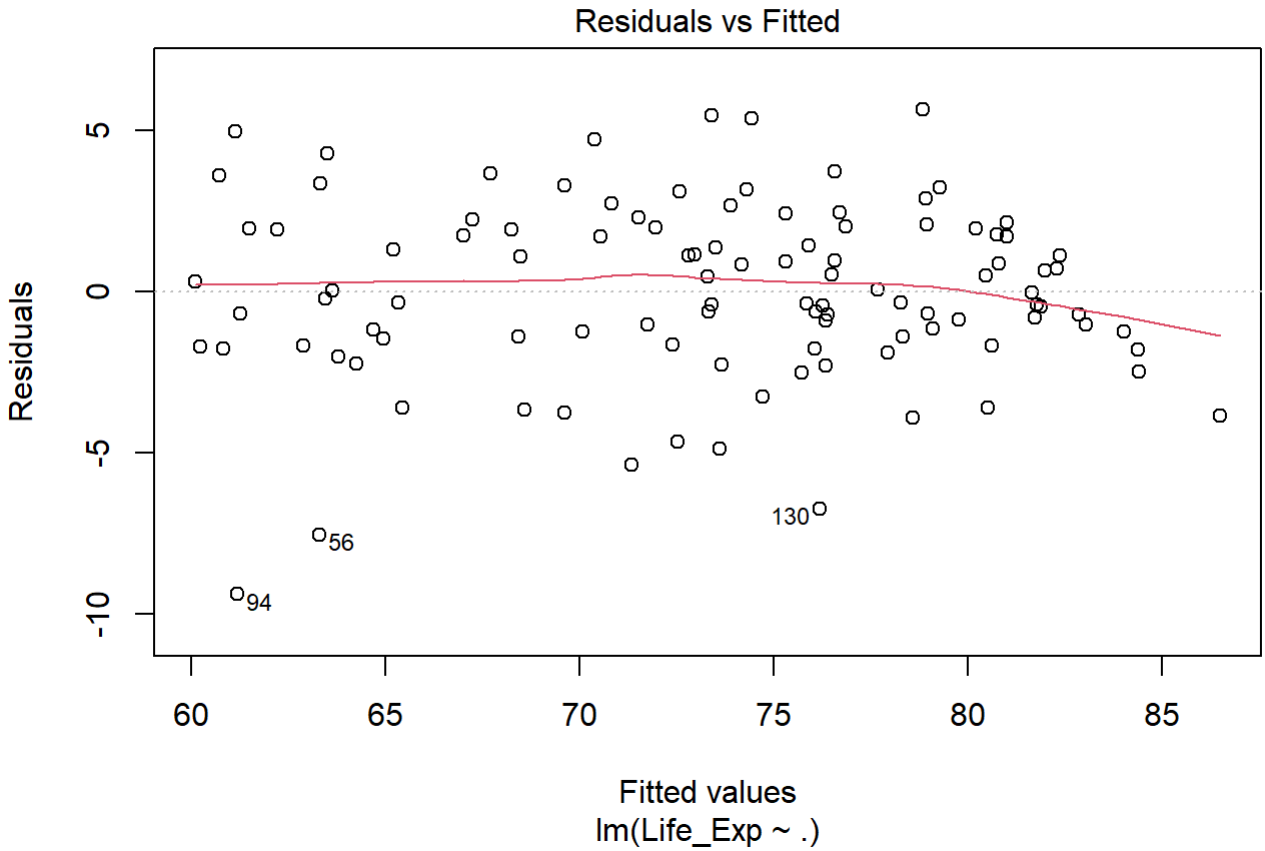

```
hist(residuals(test_mod), xlab="Residuals", main="")
```



```
shapiro.test(residuals(test_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.97821, p-value = 0.07848
```

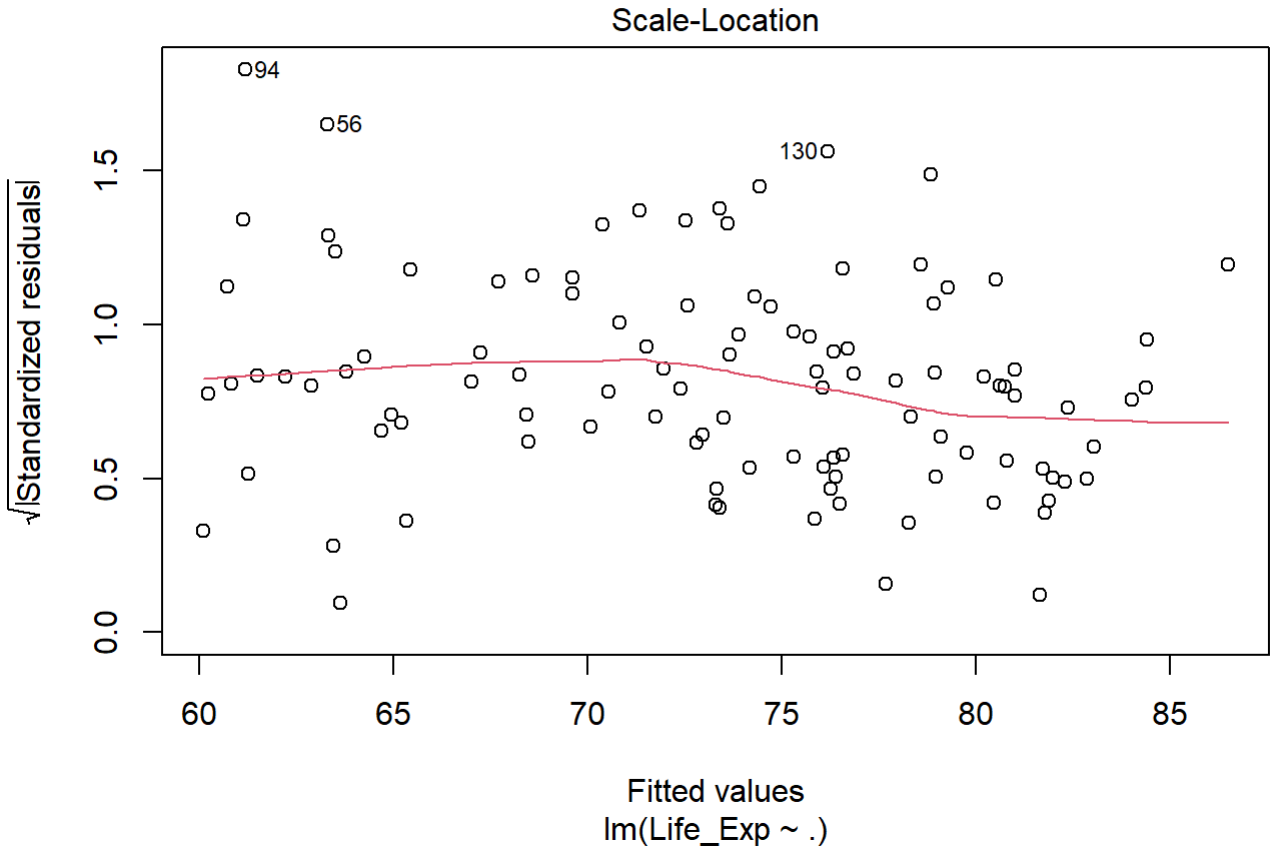
```
plot(test_mod,1)
```




```
dwtest(test_mod)
```

```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 1.8122, p-value = 0.1606
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(test_mod, 3)
```



R1 - remove highest: PM2.5

```
## AIC test:  559.4082
```

R2 - remove highest: Pol3

```
## AIC test:  557.4869
```

R3 - remove highest: Pop

```
## AIC test:  555.5546
```

R4 - remove highest: P_Ciga

```
## AIC test:  553.6612
```

R5 - remove highest: R_Bld_P

```
## AIC test:  551.8695
```

R6 - remove highest: HDL_Chol

```
## AIC test:  550.1706
```

R7 - remove highest: DTP3

```
## AIC test:  548.9232
```

R8 - remove highest: Hep83

```
## AIC test:  570.3777
```

We can observe an increase in AIC score, therefore we would add Acs_Water_Service back in model.

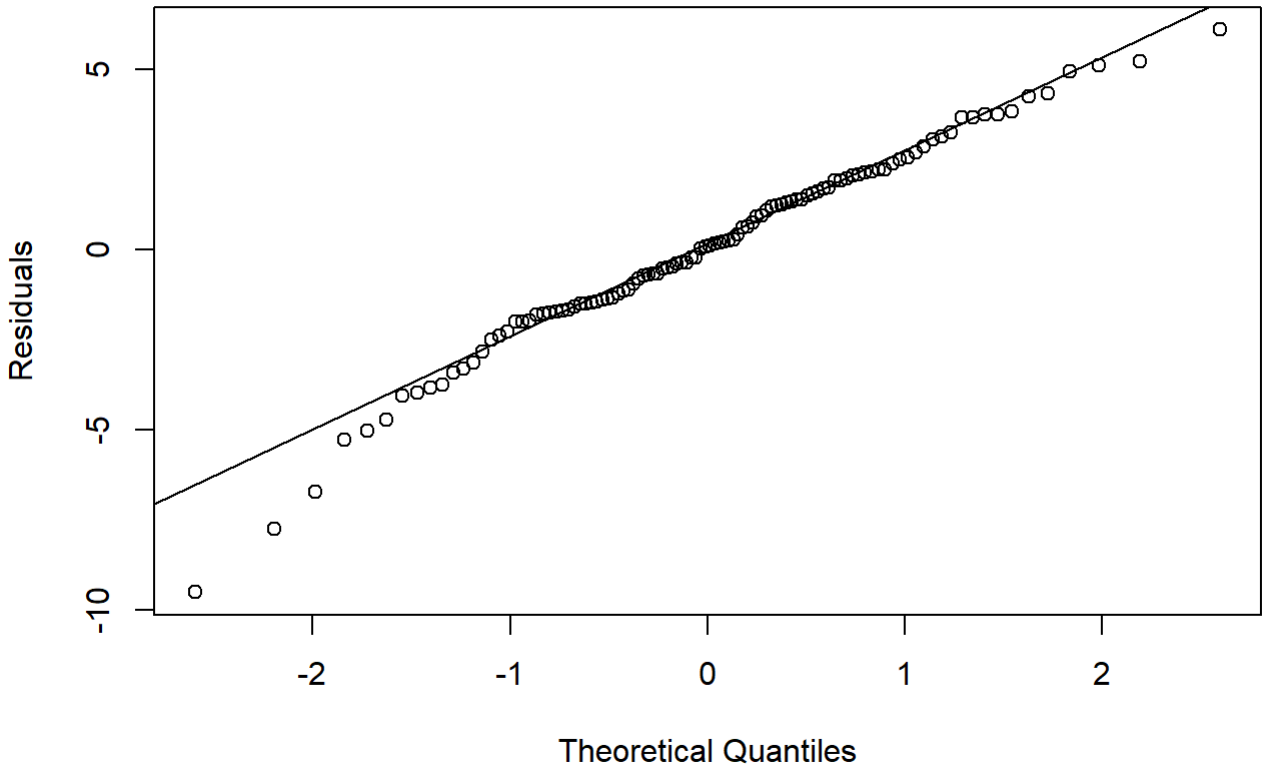
```
#Update
BE_com_lmod <- update(BE_com_lmod, . ~ . +Hep83  )


test_mod <- BE_com_lmod
summary(test_mod)
```

```
##
## Call:
## lm(formula = Life_Exp ~ Acs_Water_Service + BMI18 + BMI25 + BMI30 +
##     Diab + Hyp_Ten + Insuf_Phy + N_HDL_Chol + Alc_pcgmp + MCV1 +
##     PCGMP + EduAvg + Tax_Ciga + Cal_Daily + Hep83, data = df_184)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5045 -1.5581  0.0904  1.9251  6.0983
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.250e+01  4.872e+00   6.672    2e-09 ***
## Acs_Water_Service  6.431e-02  4.651e-02   1.383  0.17018
## BMI18              1.936e-01  1.657e-01   1.168  0.24571
## BMI25              2.657e-01  1.121e-01   2.371  0.01989 *
## BMI30             -2.566e-01  1.370e-01  -1.873  0.06426 .
## Diab              -1.821e-01  8.029e-02  -2.269  0.02568 *
## Hyp_Ten           -1.865e-01  5.536e-02  -3.368  0.00112 **
## Insuf_Phy          9.480e-02  3.457e-02   2.742  0.00737 **
## N_HDL_Chol         3.650e+00  1.190e+00   3.068  0.00284 **
## Alc_pcgmp         -2.757e-01  1.415e-01  -1.948  0.05451 .
## MCV1               4.066e-02  6.146e-02   0.662  0.50998
## PCGMP              6.613e-05  2.171e-05   3.045  0.00305 **
## EduAvg             3.831e-01  2.307e-01   1.661  0.10029
## Tax_Ciga           3.845e+00  1.944e+00   1.978  0.05102 .
## Cal_Daily          3.089e-03  1.096e-03   2.817  0.00595 **
## Hep83              2.613e-02  5.855e-02   0.446  0.65648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.98 on 90 degrees of freedom
##   (77 observations deleted due to missingness)
## Multiple R-squared:  0.8602, Adjusted R-squared:  0.8369
## F-statistic: 36.92 on 15 and 90 DF,  p-value: < 2.2e-16
```

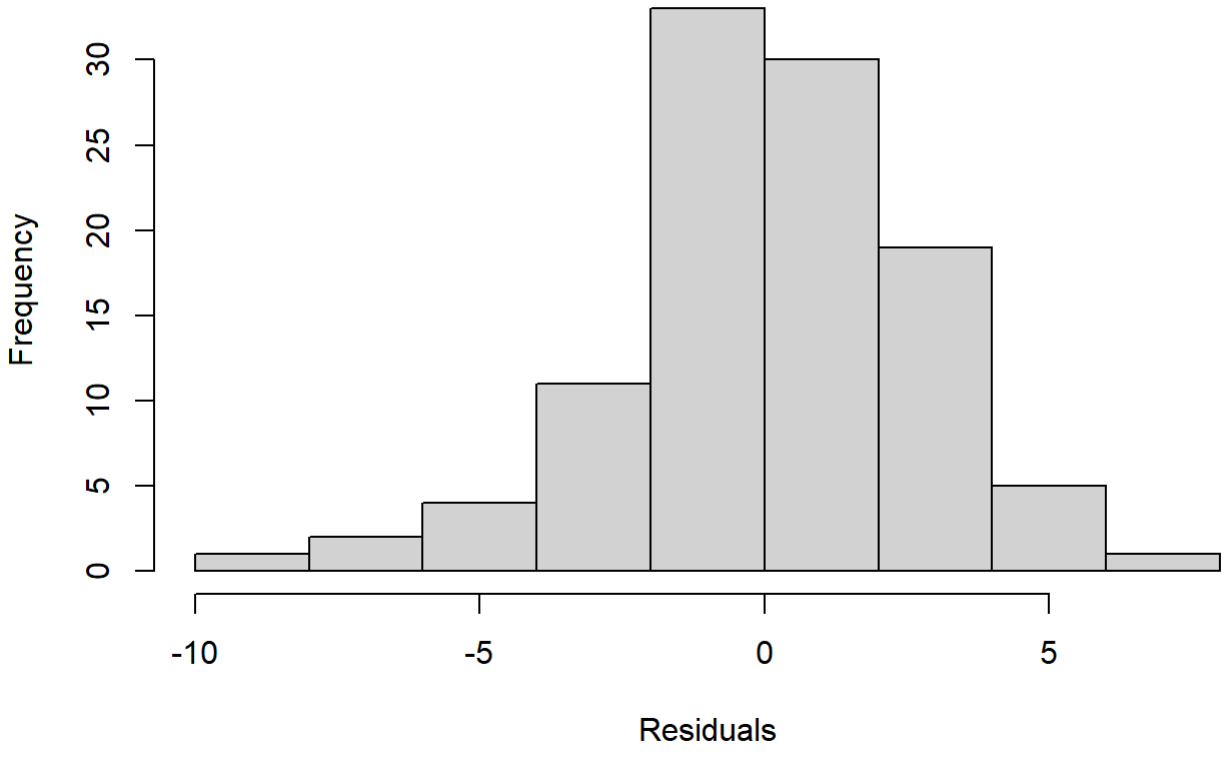
```
cat("AIC test: ", AIC(test_mod, k = 2))
```

```
## AIC test:  548.9232
```

```
qqnorm(residuals(test_mod), ylab = "Residuals", main="")
qqline(residuals(test_mod))
```



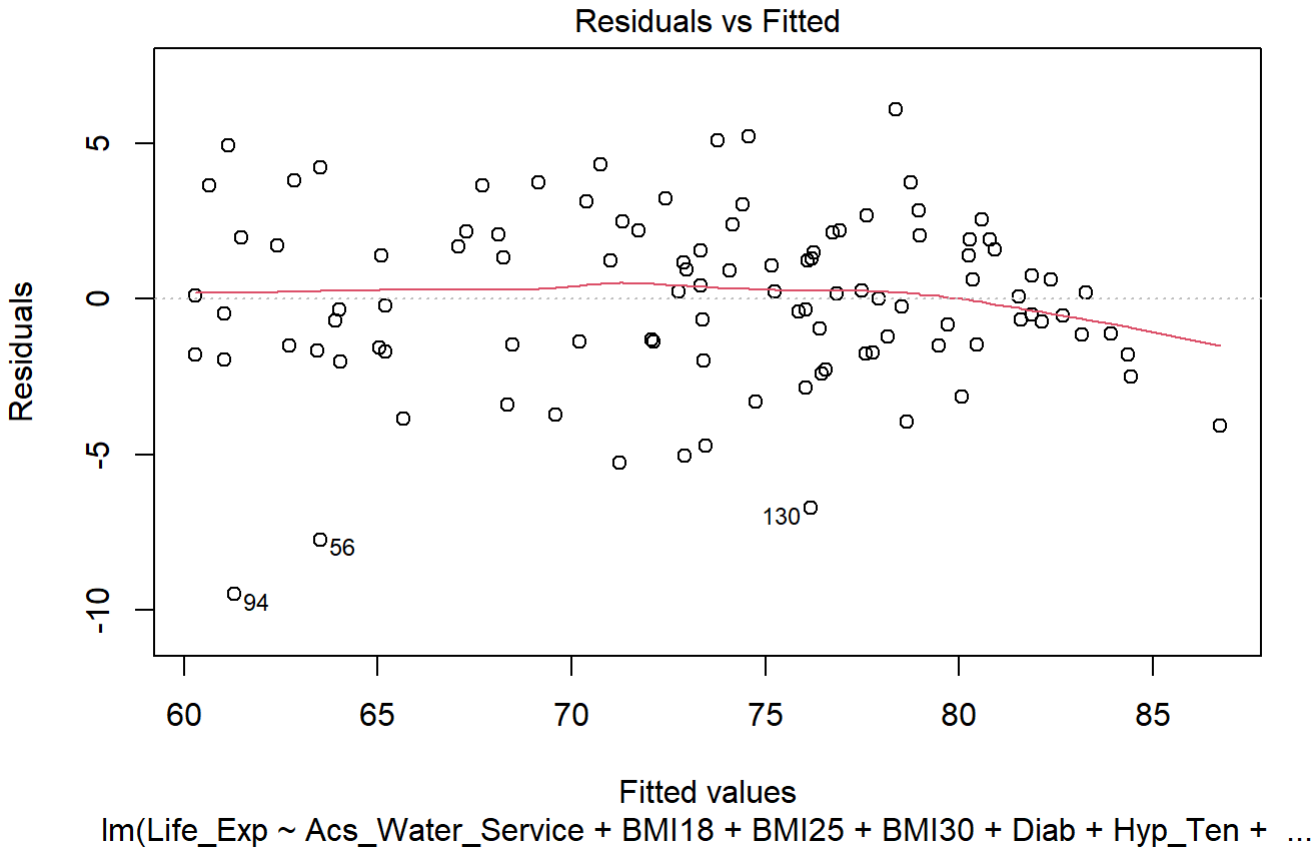

```
hist(residuals(test_mod), xlab="Residuals", main="")
```



```
shapiro.test(residuals(test_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.97965, p-value = 0.1032
```

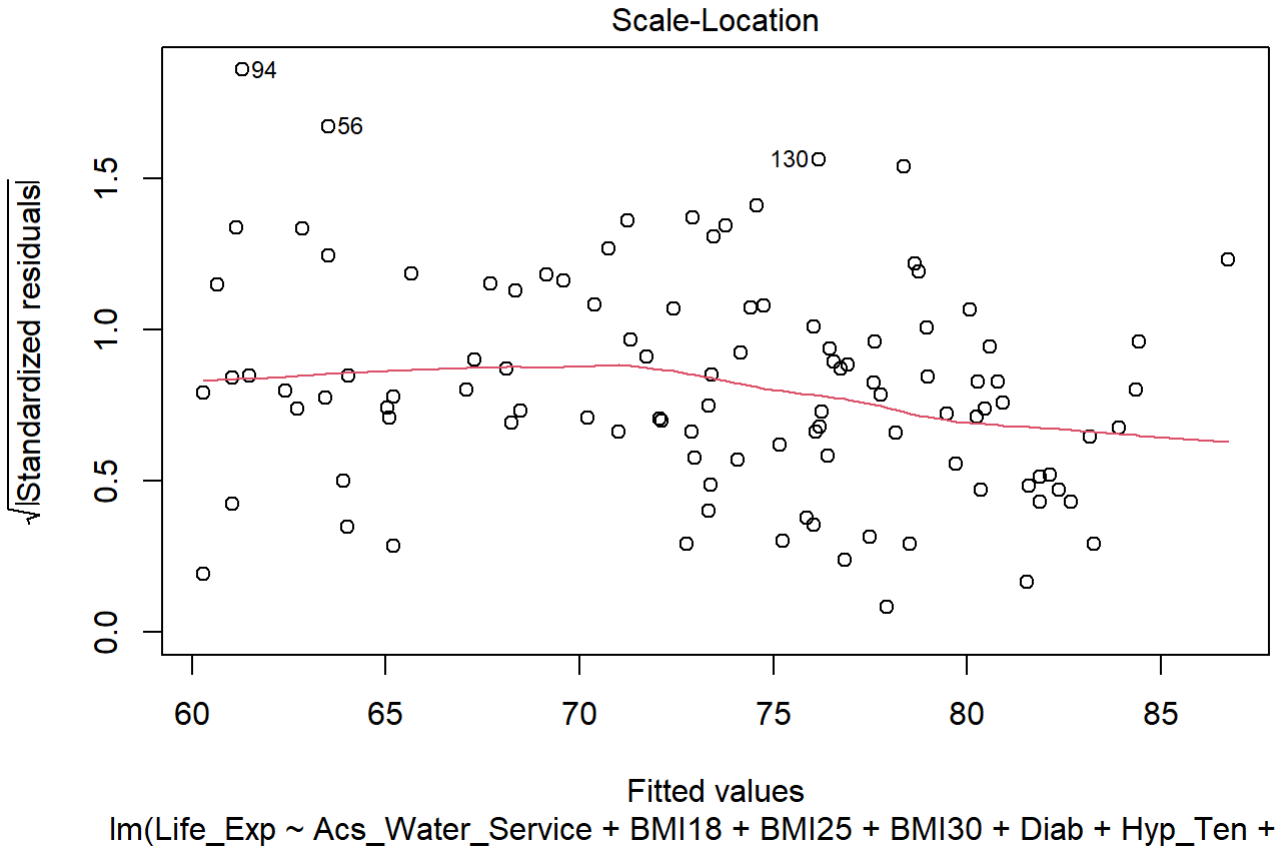
```
plot(test_mod,1)
```




```
dwtest(test_mod)
```

```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 1.7781, p-value = 0.1311
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(test_mod, 3)
```



Approach : Mean Imputation

Mean_184

```
#replacing values with mean
m_df <- as.data.frame(lapply(df_184, function(x) {
  x[is.na(x)] <- mean(x, na.rm = TRUE)
  return(x)
}))

cmeans <- colMeans(df_184, na.rm = TRUE)
cmeans
```

```
##          Life_Exp Acs_Water_Service            BMI18            BMI25
##      7.254951e+01      8.761851e+01      6.175301e+00      4.905760e+01
##            BMI30              Diab           Hyp_Ten         Insuf_Phy
##      2.074743e+01      1.262568e+01      3.736940e+01      2.669885e+01
##        N_HDL_Chol             PM2.5           R_Bld_P         Alc_pcgmp
##      3.299448e+00      2.215230e+01      3.019820e+01      4.646542e+00
##            Hep83              MCV1              Pol3              DTP3
##      8.740449e+01      8.715301e+01      8.806557e+01      8.819126e+01
##            PCGMP               Pop            EduAvg           P_Ciga
##      1.411533e+04      4.175932e+07      9.250714e+00      3.888497e+00
##         Tax_Ciga          HDL_Chol         Cal_Daily
##      5.380925e-01      1.203086e+00      2.913538e+03
```

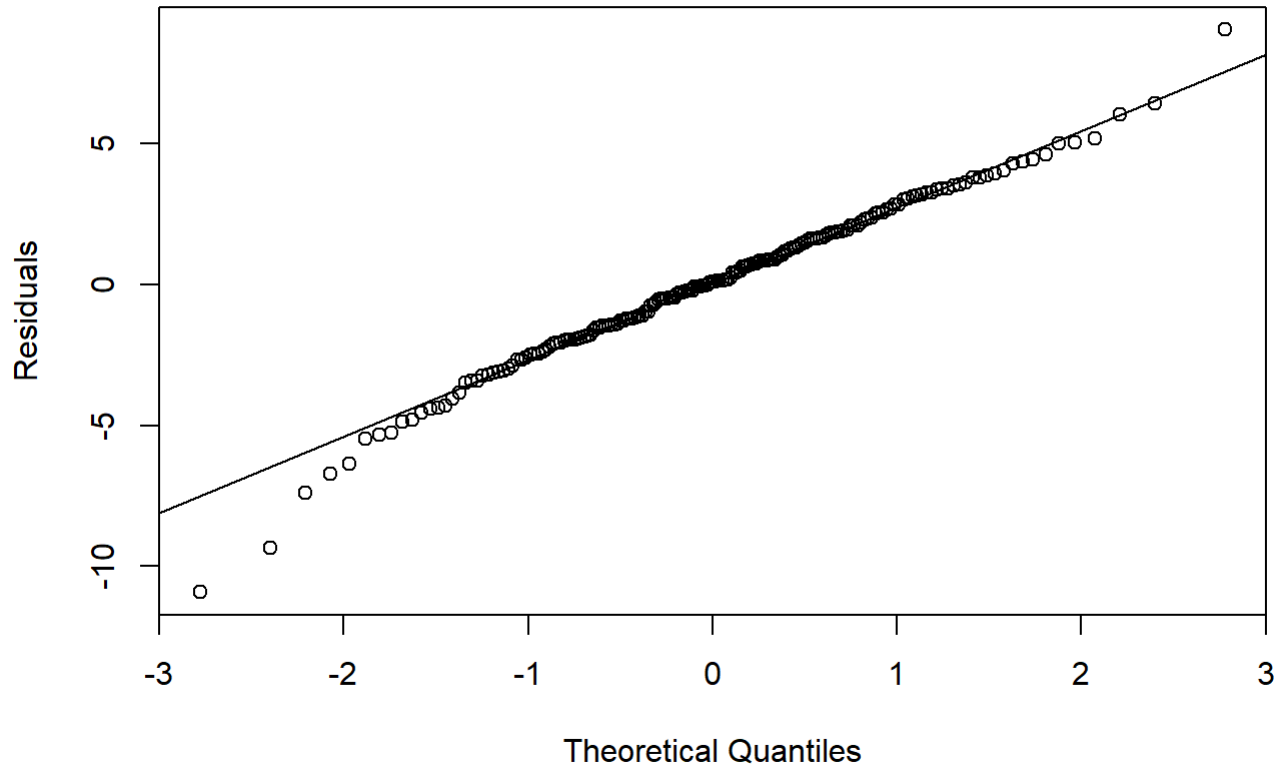
```
M_lmod <- lm(Life_Exp ~. , data= m_df)

test_mod <- M_lmod
summary(test_mod)
```

```
##
## Call:
## lm(formula = Life_Exp ~ ., data = m_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9171  -1.7910   0.1042   1.8640   9.0493
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.752e+01  4.523e+00   8.296 4.21e-14 ***
## Acs_Water_Service  1.204e-01  2.978e-02   4.042 8.21e-05 ***
## BMI18             -8.234e-02  1.040e-01  -0.792 0.429662
## BMI25              1.652e-01  7.495e-02   2.204 0.028946 *
## BMI30             -2.454e-01  7.927e-02  -3.096 0.002317 **
## Diab              -1.330e-01  6.470e-02  -2.056 0.041440 *
## Hyp_Ten           -1.125e-01  1.144e-01  -0.984 0.326761
## Insuf_Phy          1.117e-01  2.986e-02   3.740 0.000256 ***
## N_HDL_Chol         3.156e+00  9.843e-01   3.206 0.001624 **
## PM2.5              1.374e-02  2.380e-02   0.577 0.564593
## R_Bld_P           -3.852e-03  1.051e-01  -0.037 0.970810
## Alc_pcgmp          3.132e-02  1.007e-01   0.311 0.756293
## Hep83              7.598e-03  7.258e-02   0.105 0.916754
## MCV1               8.119e-03  3.812e-02   0.213 0.831602
## Pol3               1.267e-01  9.850e-02   1.286 0.200201
## DTP3              -6.731e-02  1.139e-01  -0.591 0.555237
## PCGMP              6.119e-05  2.334e-05   2.621 0.009603 **
## Pop               -3.244e-10  1.704e-09  -0.190 0.849258
## EduAvg             1.902e-01  1.712e-01   1.111 0.268266
## P_Ciga             1.080e-01  1.076e-01   1.004 0.317034
## Tax_Ciga           3.639e+00  1.297e+00   2.806 0.005647 **
## HDL_Chol          -1.396e+00  2.766e+00  -0.505 0.614502
## Cal_Daily          1.452e-03  8.960e-04   1.621 0.107001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.081 on 160 degrees of freedom
## Multiple R-squared:  0.8428, Adjusted R-squared:  0.8212
## F-statistic: 38.99 on 22 and 160 DF,  p-value: < 2.2e-16
```

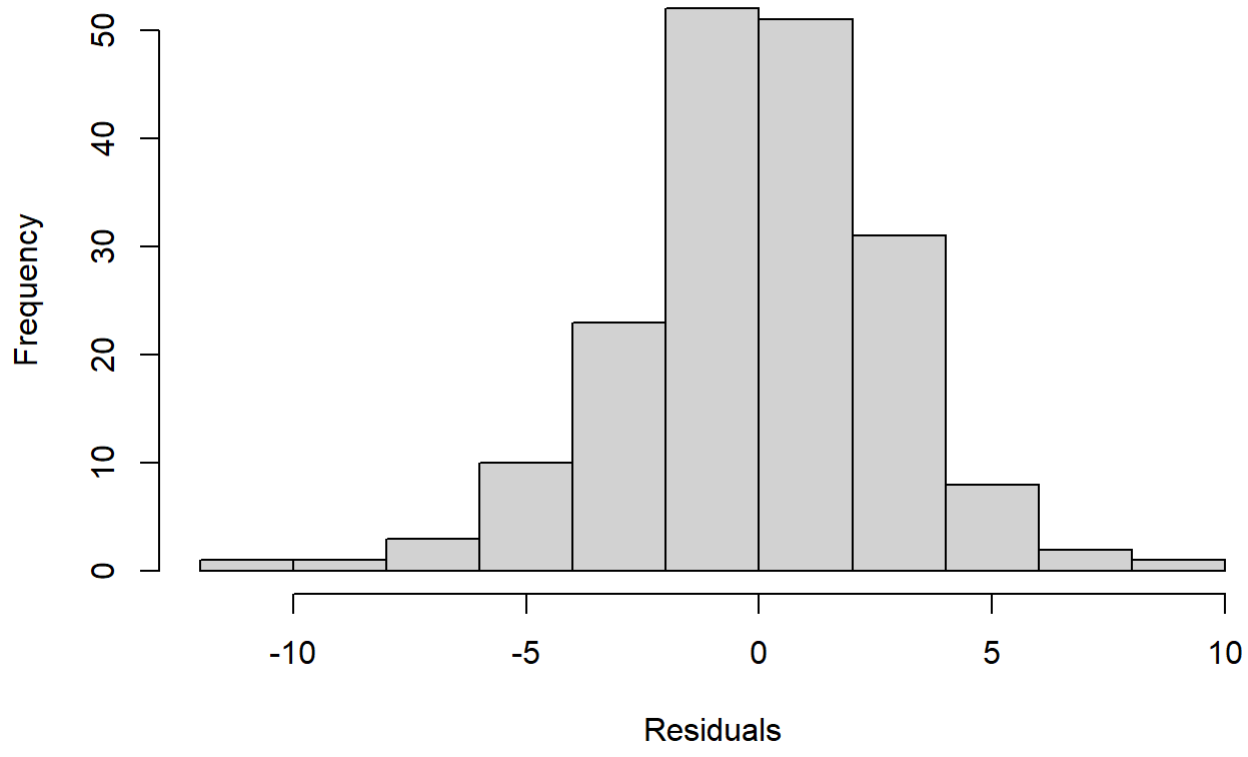
```
cat("AIC test: ", AIC(test_mod, k = 2))
```

```
## AIC test:  954.587
```

```
qqnorm(residuals(test_mod), ylab = "Residuals", main="")
qqline(residuals(test_mod))
```



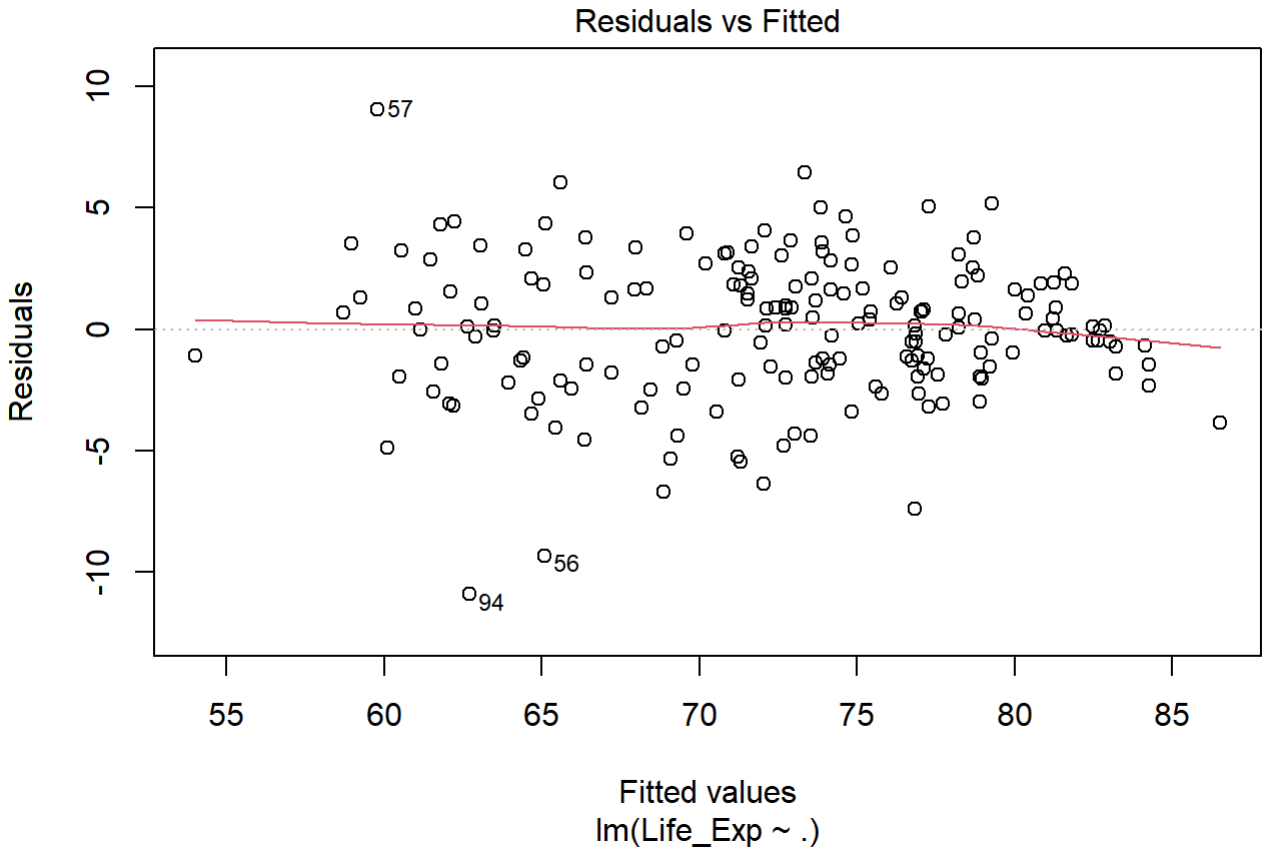

```
hist(residuals(test_mod), xlab="Residuals", main="")
```



```
shapiro.test(residuals(test_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.98527, p-value = 0.05191
```

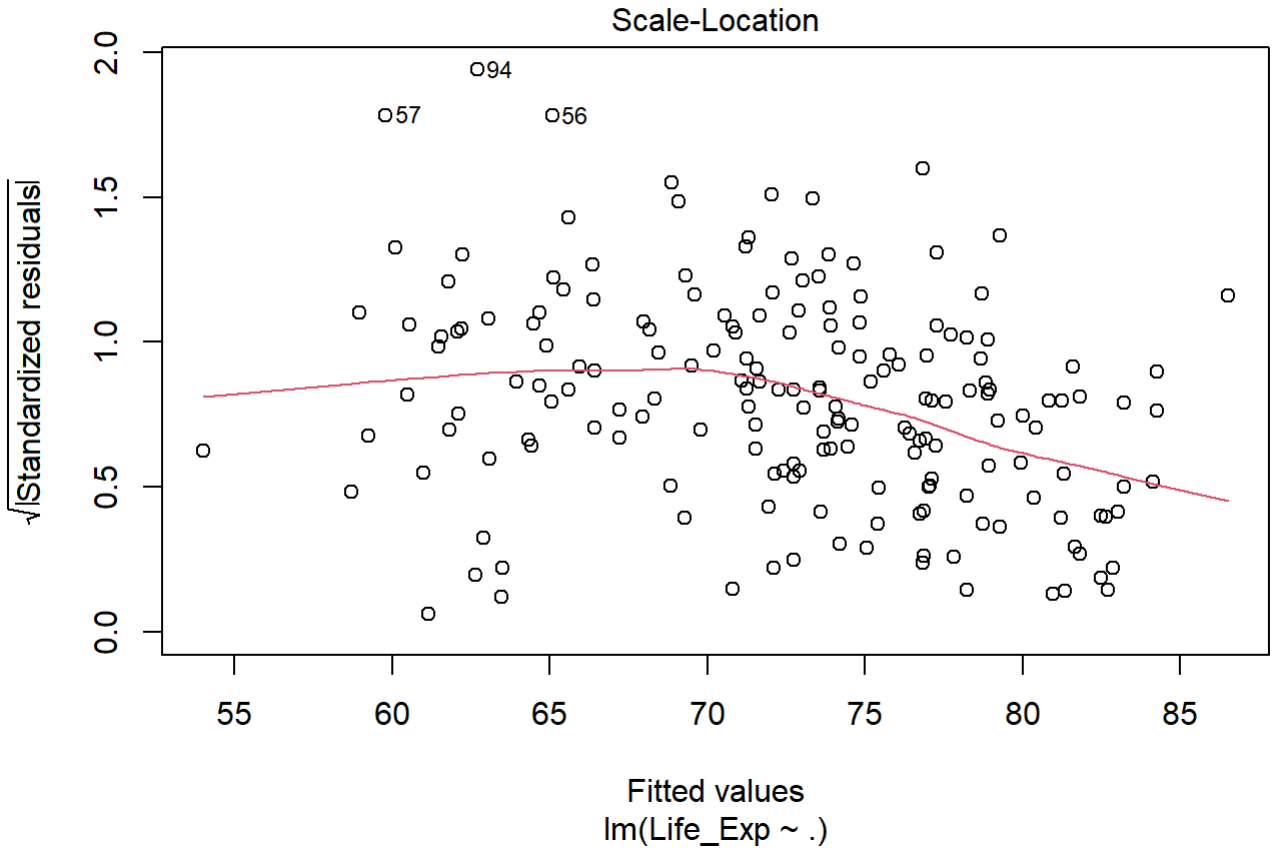
```
plot(test_mod,1)
```




```
dwtest(test_mod)
```

```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 2.0605, p-value = 0.6599
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(test_mod, 3)
```



R1 - remove highest: R_Bld_P

```
## AIC test:  952.5885
```

R2 - remove highest: Hep83

```
## AIC test:  950.6005
```

R3 - remove highest: Pop

```
## AIC test:  948.6471
```

R4 - remove highest: MCV1

```
## AIC test:  946.7159
```

R5 - remove highest: Alc_pcgmp

```
## AIC test:  944.835
```

R6 - remove highest: HDL_Chol

```
## AIC test:  943.076
```

R7 - remove highest: PM2.5

```
## AIC test:  941.2992
```

R8 - remove highest: DTP3

```
## AIC test:  939.6415
```

R9 - remove highest: BMI18

```
## AIC test:  938.275
```

R10 - remove highest: P_Ciga

```
## AIC test:  937.0903
```

R11 - remove highest: EduAvg

```
## AIC test:  936.4524
```

R12 - remove highest: Cal_Daily

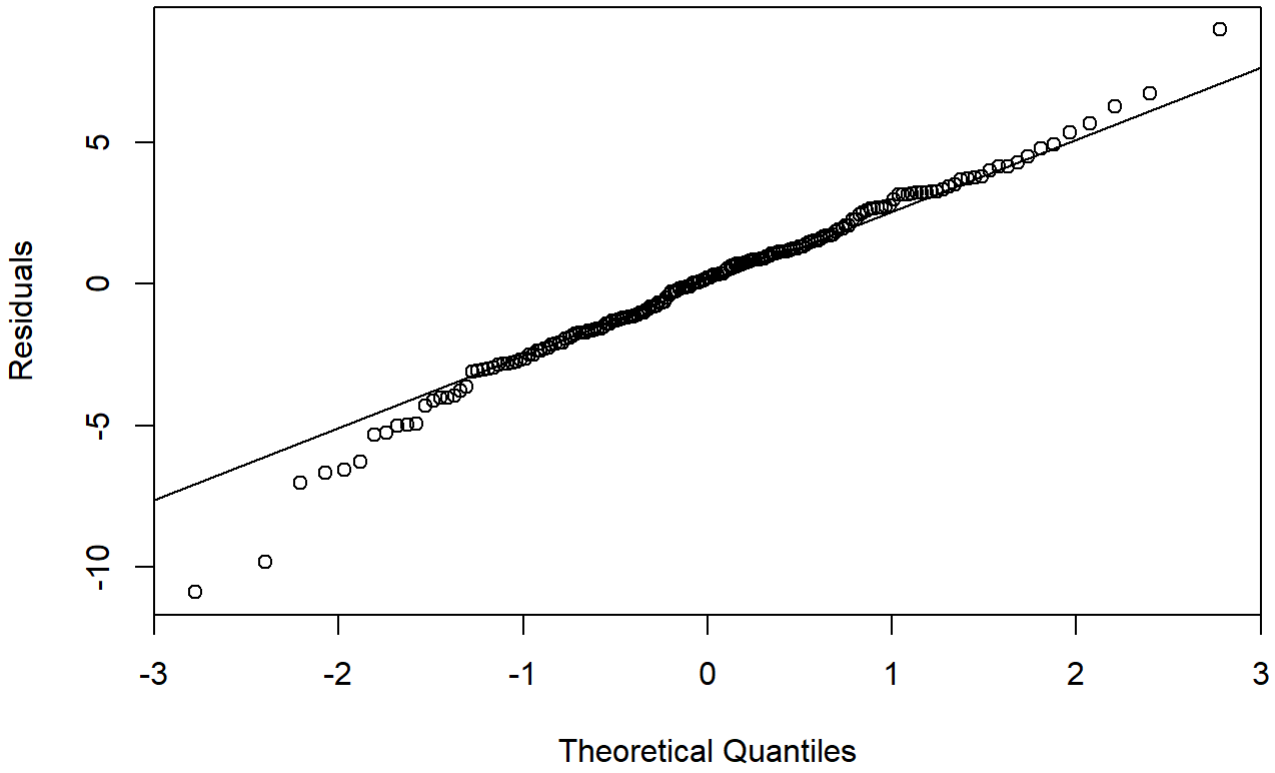
```
## AIC test:  938.291
```

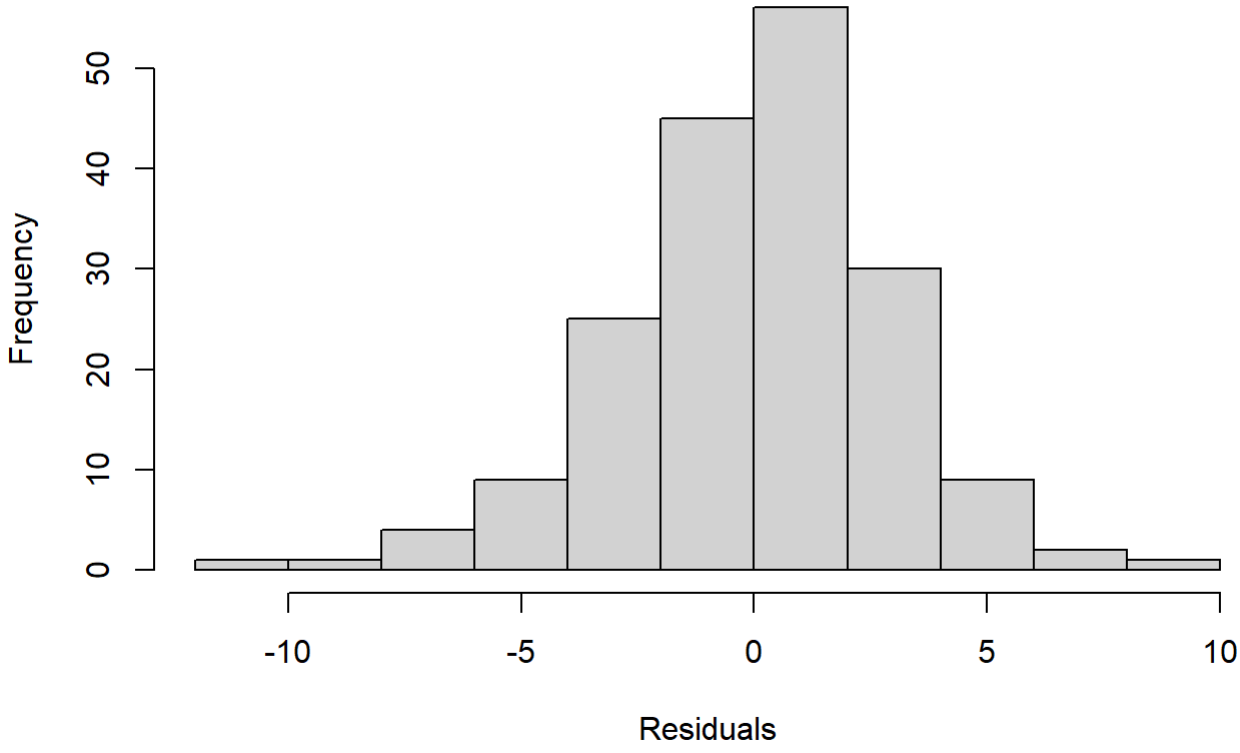
AIC increases, so we add back Cal_Daily

Final Result of Approach 2

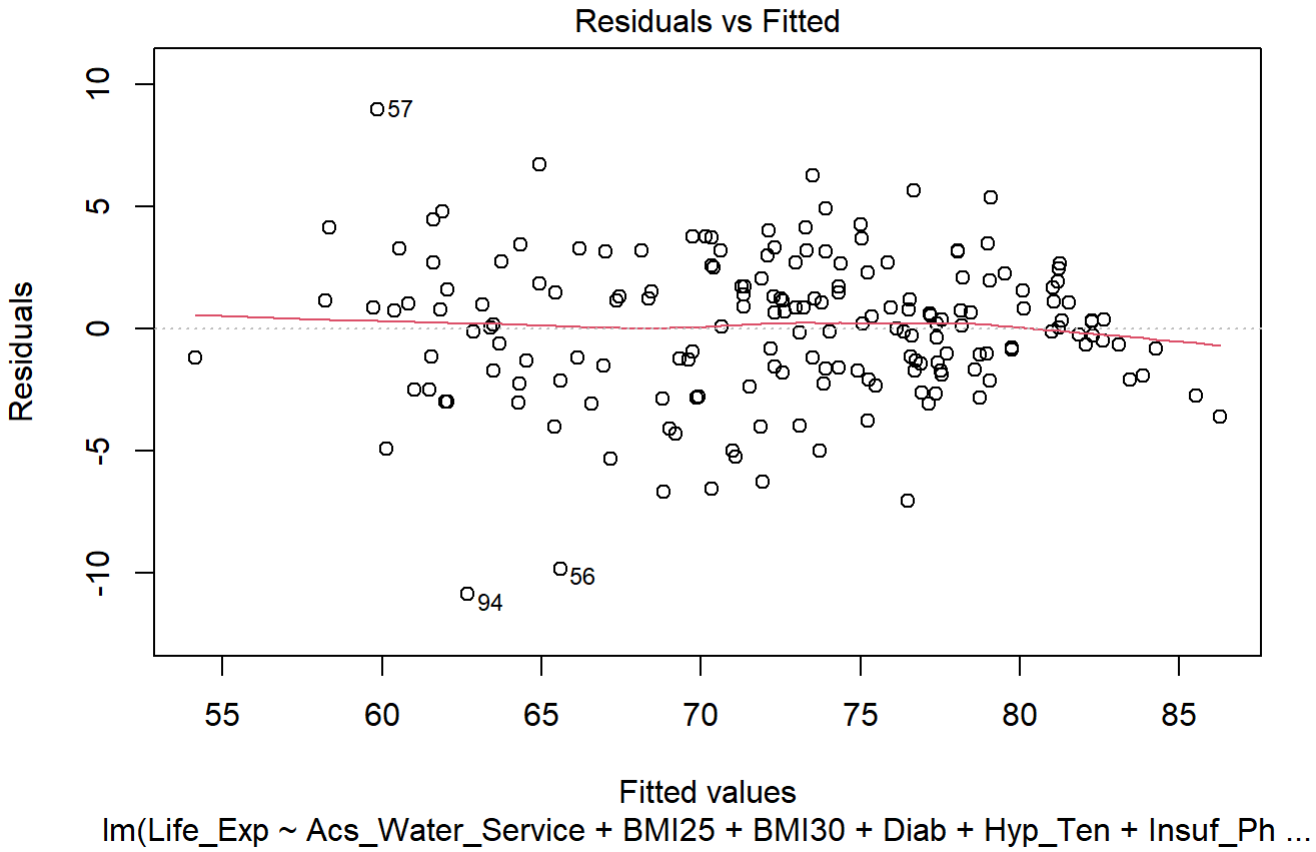
```
##
## Call:
## lm(formula = Life_Exp ~ Acs_Water_Service + BMI25 + BMI30 + Diab +
##     Hyp_Ten + Insuf_Phy + N_HDL_Chol + Pol3 + PCGMP + Tax_Ciga +
##     Cal_Daily, data = m_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8964  -1.7093   0.2057   1.7265   8.9824
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.567e+01  3.187e+00  11.191  < 2e-16 ***
## Acs_Water_Service  1.262e-01  2.742e-02   4.605 8.03e-06 ***
## BMI25              2.047e-01  5.261e-02   3.891 0.000143 ***
## BMI30             -2.547e-01  6.879e-02  -3.703 0.000287 ***
## Diab              -1.429e-01  5.301e-02  -2.696 0.007730 **
## Hyp_Ten           -1.270e-01  3.862e-02  -3.288 0.001224 **
## Insuf_Phy          1.064e-01  2.372e-02   4.487 1.32e-05 ***
## N_HDL_Chol         3.119e+00  8.686e-01   3.591 0.000430 ***
## Pol3               7.625e-02  2.260e-02   3.375 0.000915 ***
## PCGMP              7.411e-05  1.643e-05   4.511 1.20e-05 ***
## Tax_Ciga           3.888e+00  1.214e+00   3.202 0.001628 **
## Cal_Daily          1.587e-03  8.337e-04   1.904 0.058604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.012 on 171 degrees of freedom
## Multiple R-squared:  0.8394, Adjusted R-squared:  0.8291
## F-statistic: 81.28 on 11 and 171 DF,  p-value: < 2.2e-16
```

```
## AIC test:  936.4524
```

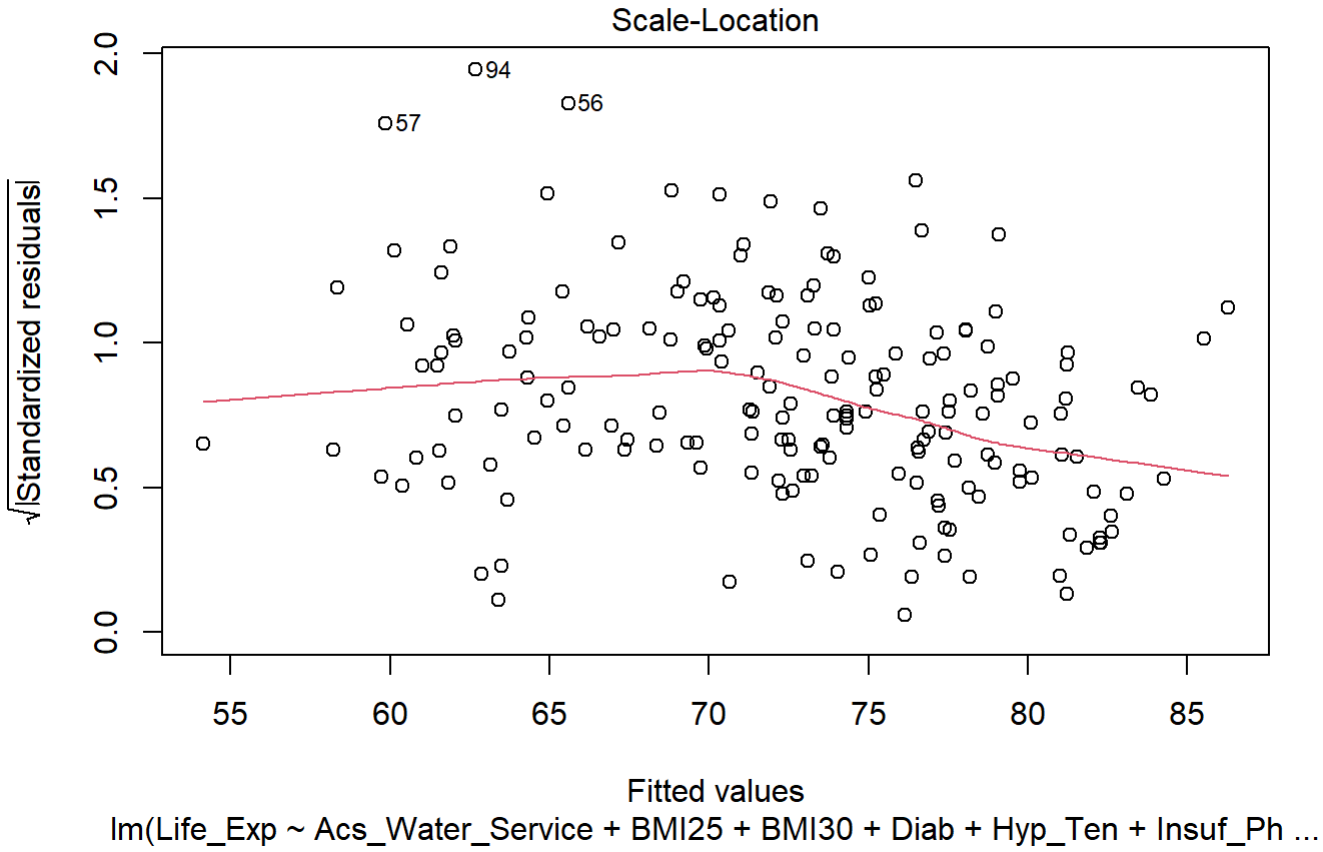

Residuals
5
0
-5
-10
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Frequency
0
10
20
30
40
50
-10
-5
0
5
10
Residuals

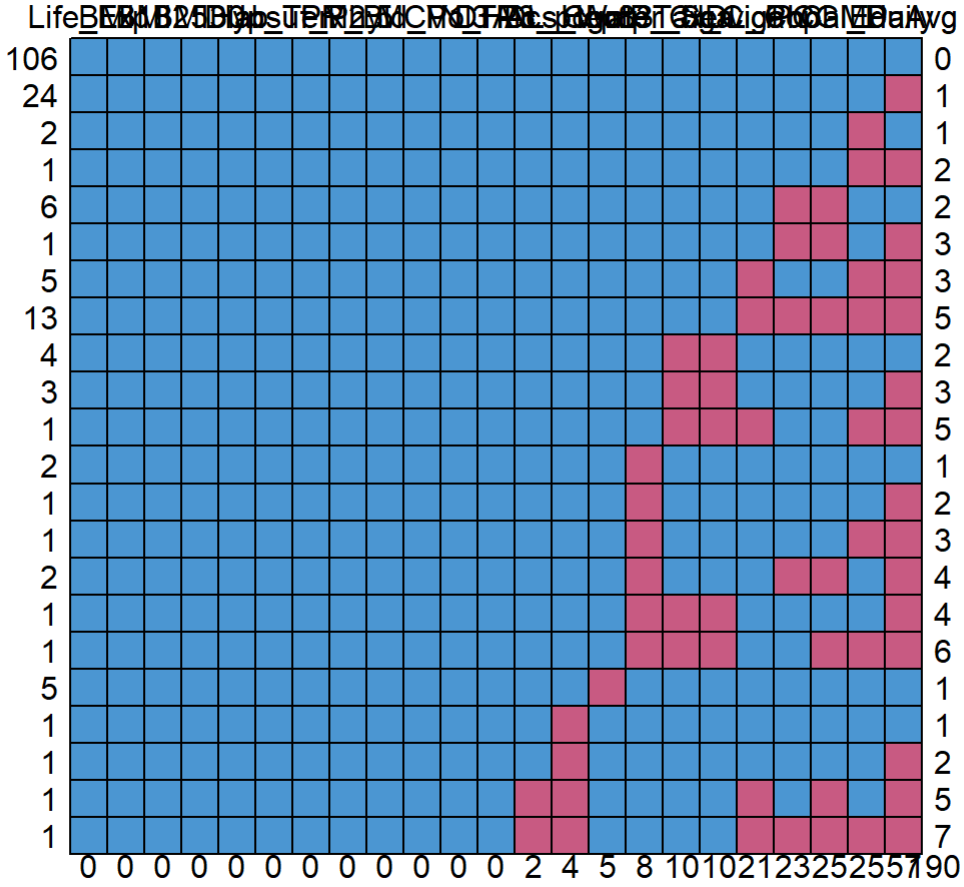
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.98321, p-value = 0.02701
```




```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 1.9954, p-value = 0.4917
## alternative hypothesis: true autocorrelation is greater than 0
```



Approach 3: Imputation by multiple regression on missing value


106 0
24 1
2 1
1 2
6 2
1 3
5 3
13 5
4 2
3 3
1 5
2 1
1 2
1 3
2 4
1 4
1 6
5 1
1 1
1 2
1 5
1 7
0 0 0 0 0 0 0 0 0 0 0 0 2 4 5 8 10 10 21 23 25 25 57 90

```
##     Life_Exp BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy PM2.5 R_Bld_P MCV1 Pol3
## 106        1     1     1     1    1       1         1     1       1    1    1
## 24         1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 6          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 5          1     1     1     1    1       1         1     1       1    1    1
## 13         1     1     1     1    1       1         1     1       1    1    1
## 4          1     1     1     1    1       1         1     1       1    1    1
## 3          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 5          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
##            0     0     0     0    0       0         0     0       0    0    0
##     DTP3 N_HDL_Chol Alc_pcgmp Hep83 Acs_Water_Service P_Ciga Tax_Ciga HDL_Chol
## 106    1          1         1     1                 1      1        1        1
## 24     1          1         1     1                 1      1        1        1
## 2      1          1         1     1                 1      1        1        1
## 1      1          1         1     1                 1      1        1        1
## 6      1          1         1     1                 1      1        1        1
## 1      1          1         1     1                 1      1        1        1
## 5      1          1         1     1                 1      1        1        0
## 13     1          1         1     1                 1      1        1        0
## 4      1          1         1     1                 1      0        0        1
## 3      1          1         1     1                 1      0        0        1
## 1      1          1         1     1                 1      0        0        0
## 2      1          1         1     1                 0      1        1        1
## 1      1          1         1     1                 0      1        1        1
## 1      1          1         1     1                 0      1        1        1
## 2      1          1         1     1                 0      1        1        1
## 1      1          1         1     1                 0      0        0        1
## 1      1          1         1     1                 0      0        0        1
## 5      1          1         1     0                 1      1        1        1
## 1      1          1         0     1                 1      1        1        1
## 1      1          1         0     1                 1      1        1        1
## 1      1          0         0     1                 1      1        1        0
## 1      1          0         0     1                 1      1        1        0
##        0          2         4     5                 8     10       10       21
##     Pop PCGMP Cal_Daily EduAvg
## 106   1     1         1      1   0
## 24    1     1         1      0   1
## 2     1     1         0      1   1
## 1     1     1         0      0   2
## 6     0     0         1      1   2
## 1     0     0         1      0   3
```

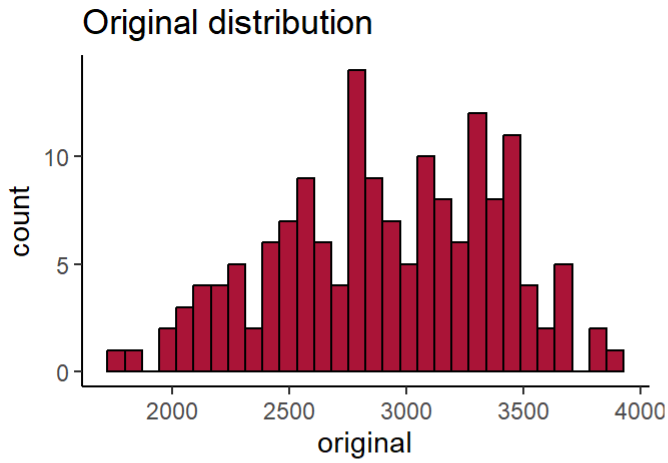
```
## 5     1     1         0      0   3
## 13    0     0         0      0   5
## 4     1     1         1      1   2
## 3     1     1         1      0   3
## 1     1     1         0      0   5
## 2     1     1         1      1   1
## 1     1     1         1      0   2
## 1     1     1         0      0   3
## 2     0     0         1      0   4
## 1     1     1         1      0   4
## 1     1     0         0      0   6
## 5     1     1         1      1   1
## 1     1     1         1      1   1
## 1     1     1         1      0   2
## 1     1     0         1      0   5
## 1     0     0         0      0   7
##      23    25        25     57 190
```

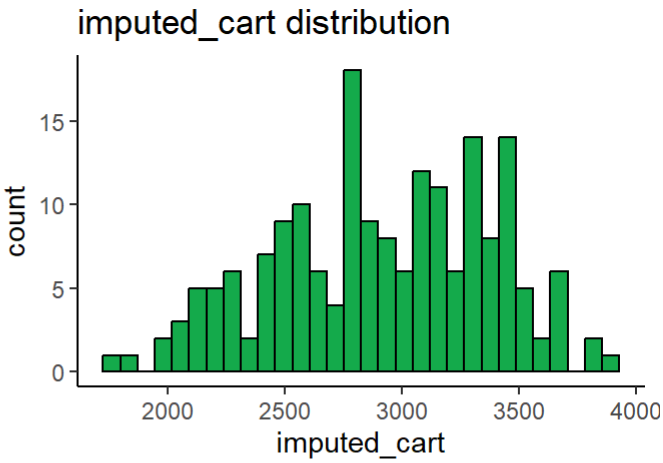
1) imputation on Cal_Daily

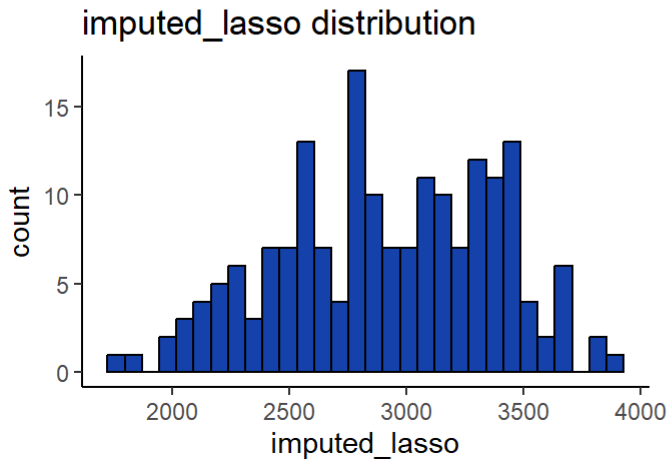
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 25 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```







From above, imputed_cart would be chosen as imputation.

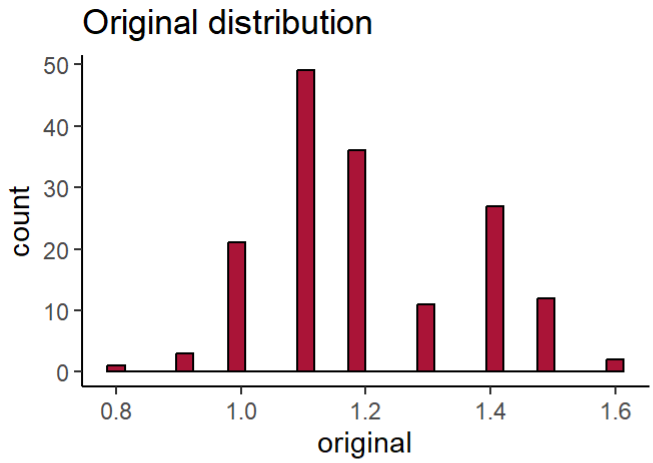
```
#replace with imputed data: Cal_Daily

Re_imp_df184 <- df_184
Re_imp_df184$Cal_Daily <- mice_imputed$imputed_cart
```

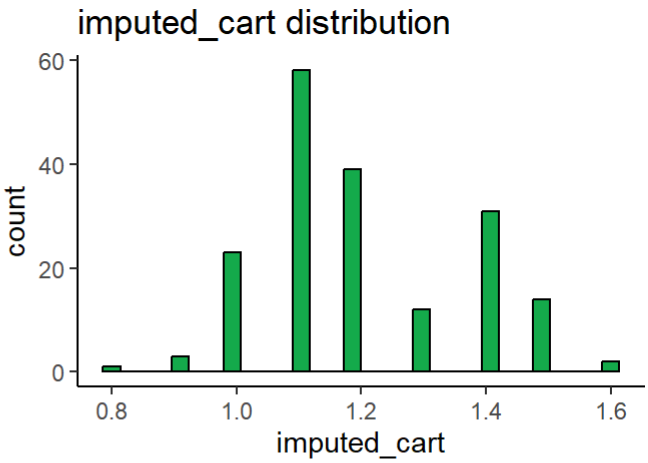
2) imputation on HDL_Chol

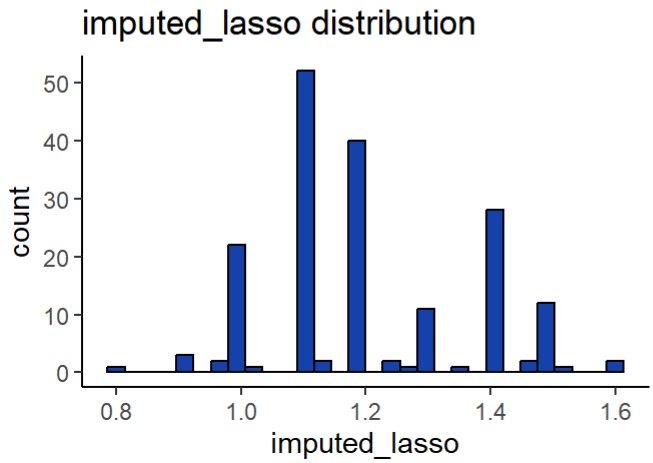
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 21 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```







From above, imputed_cart would be chosen as imputation.

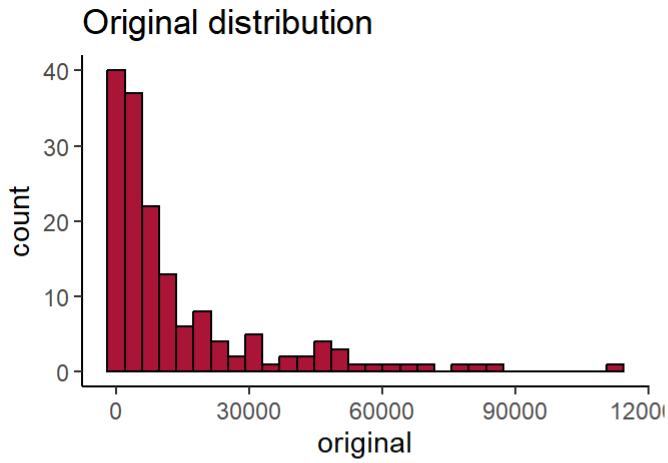
```
#replace with imputed data: HDL_Chol

Re_imp_df184$HDL_Chol <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

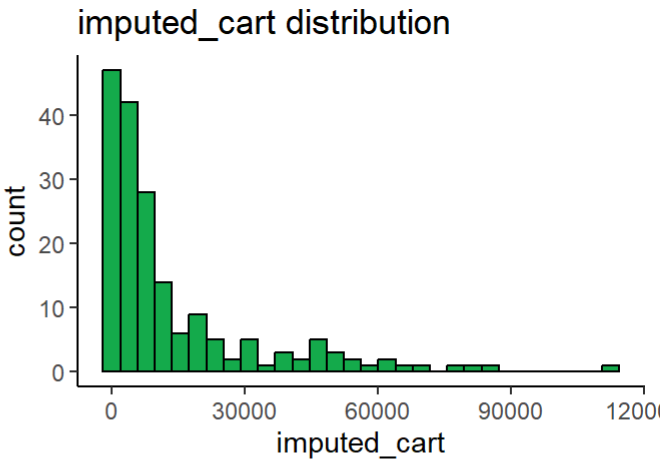
3) imputation on PCGMP

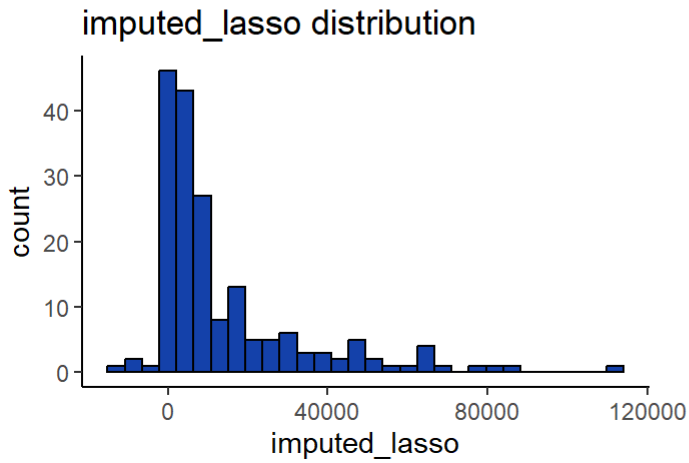
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 25 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

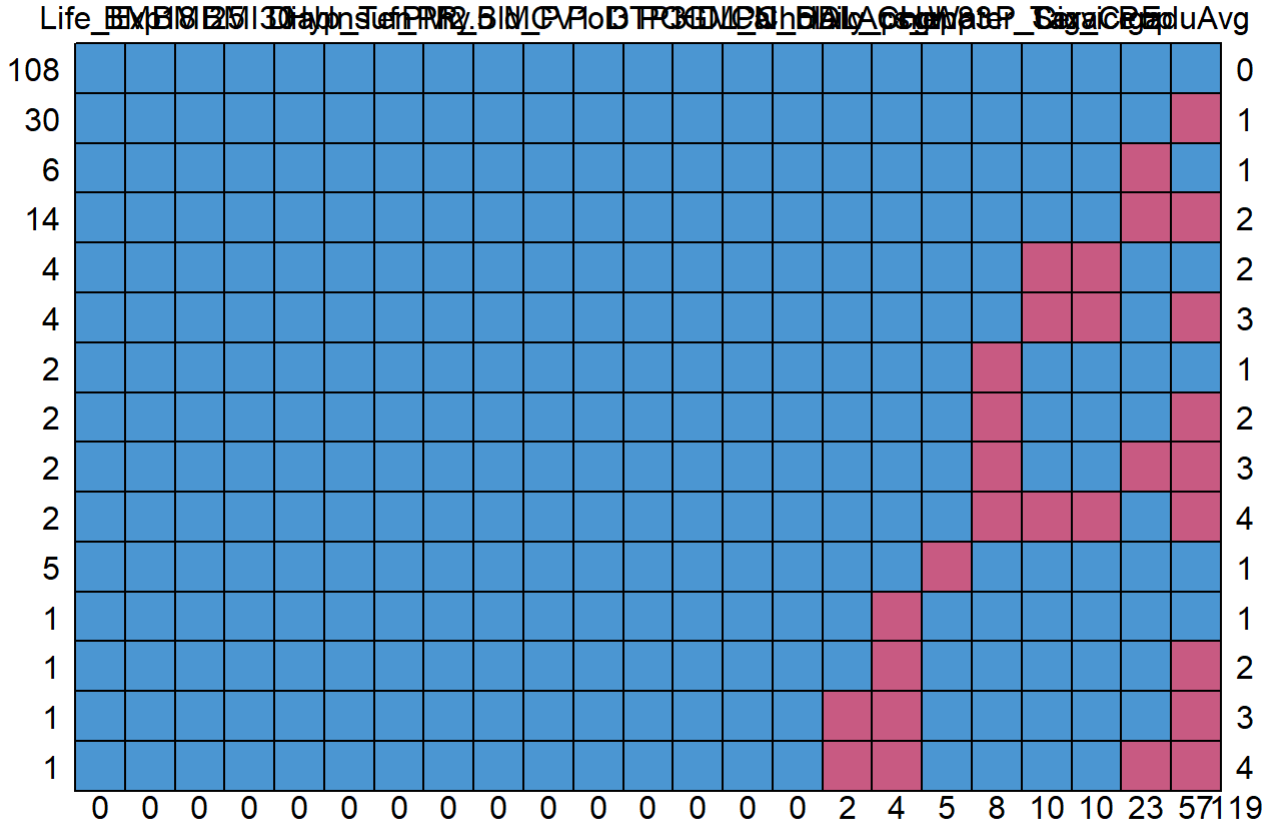






From above, imputed_cart would be chosen as imputation.

```
#replace with imputed data: PCGMP

Re_imp_df184$PCGMP <- mice_imputed$imputed_cart
md.pattern(Re_imp_df184)
```


Life_Edu
108 30 6 14 4 4 2 2 2 2 5 1 1 1
0 1 1 2 2 3 1 2 3 4 1 1 2 3 4
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 4 5 8 10 10 23 57 119

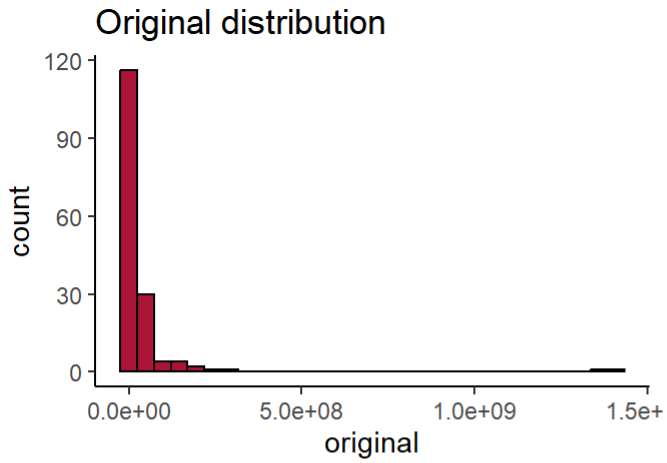
```
##     Life_Exp BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy PM2.5 R_Bld_P MCV1 Pol3
## 108        1     1     1     1    1       1         1     1       1    1    1
## 30         1     1     1     1    1       1         1     1       1    1    1
## 6          1     1     1     1    1       1         1     1       1    1    1
## 14         1     1     1     1    1       1         1     1       1    1    1
## 4          1     1     1     1    1       1         1     1       1    1    1
## 4          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 5          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
##            0     0     0     0    0       0         0     0       0    0    0
##     DTP3 PCGMP HDL_Chol Cal_Daily N_HDL_Chol Alc_pcgmp Hep83 Acs_Water_Service
## 108    1     1        1         1          1         1     1                 1
## 30     1     1        1         1          1         1     1                 1
## 6      1     1        1         1          1         1     1                 1
## 14     1     1        1         1          1         1     1                 1
## 4      1     1        1         1          1         1     1                 1
## 4      1     1        1         1          1         1     1                 1
## 2      1     1        1         1          1         1     1                 0
## 2      1     1        1         1          1         1     1                 0
## 2      1     1        1         1          1         1     1                 0
## 2      1     1        1         1          1         1     1                 0
## 5      1     1        1         1          1         1     0                 1
## 1      1     1        1         1          1         0     1                 1
## 1      1     1        1         1          1         0     1                 1
## 1      1     1        1         1          0         0     1                 1
## 1      1     1        1         1          0         0     1                 1
##        0     0        0         0          2         4     5                 8
##     P_Ciga Tax_Ciga Pop EduAvg
## 108      1        1   1      1   0
## 30       1        1   1      0   1
## 6        1        1   0      1   1
## 14       1        1   0      0   2
## 4        0        0   1      1   2
## 4        0        0   1      0   3
## 2        1        1   1      1   1
## 2        1        1   1      0   2
## 2        1        1   0      0   3
## 2        0        0   1      0   4
## 5        1        1   1      1   1
## 1        1        1   1      1   1
## 1        1        1   1      0   2
## 1        1        1   1      0   3
## 1        1        1   0      0   4
##         10       10  23     57 119
```

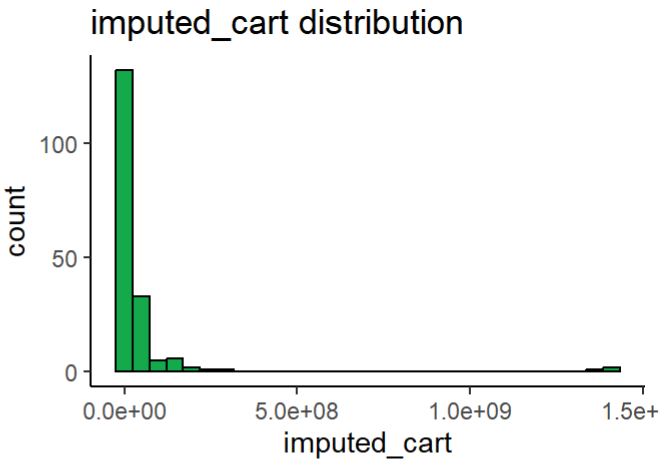
4) imputation on Pop

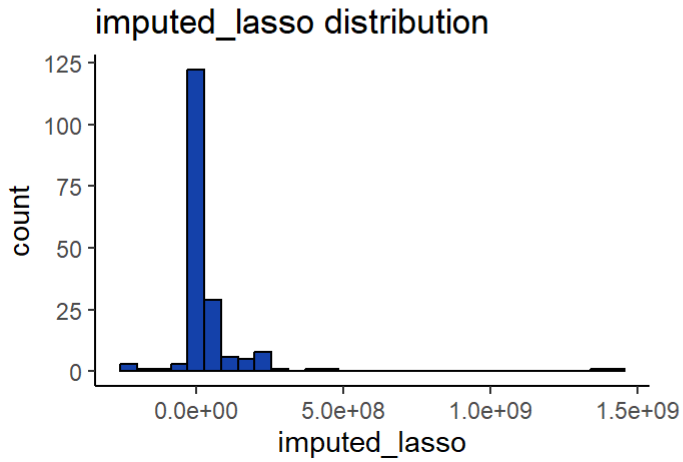
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 23 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```







From above, imputed_cart would be chosen as imputation.

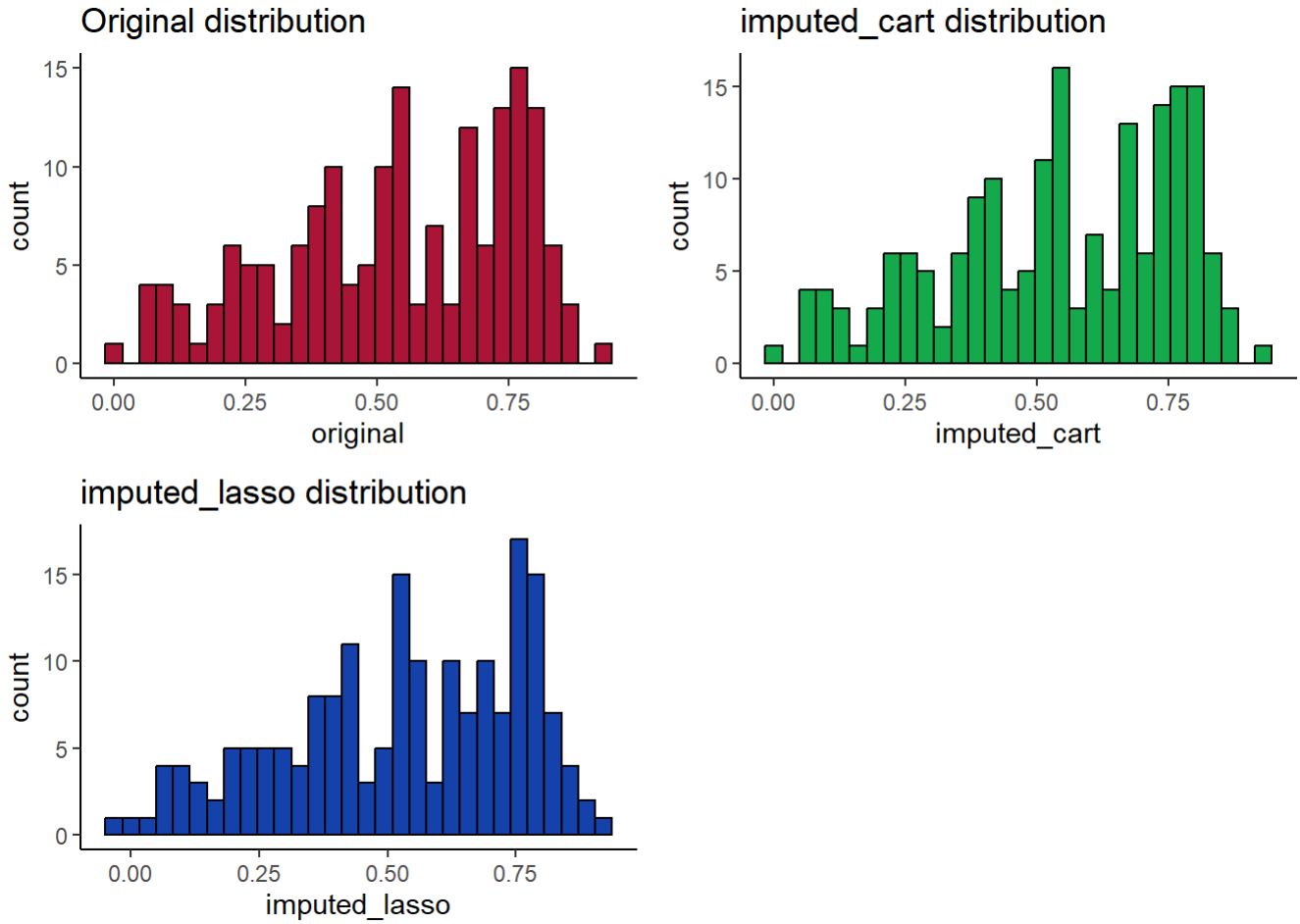
```
#replace with imputed data: Pop

Re_imp_df184$Pop <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

5) imputation on Tax_Ciga

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

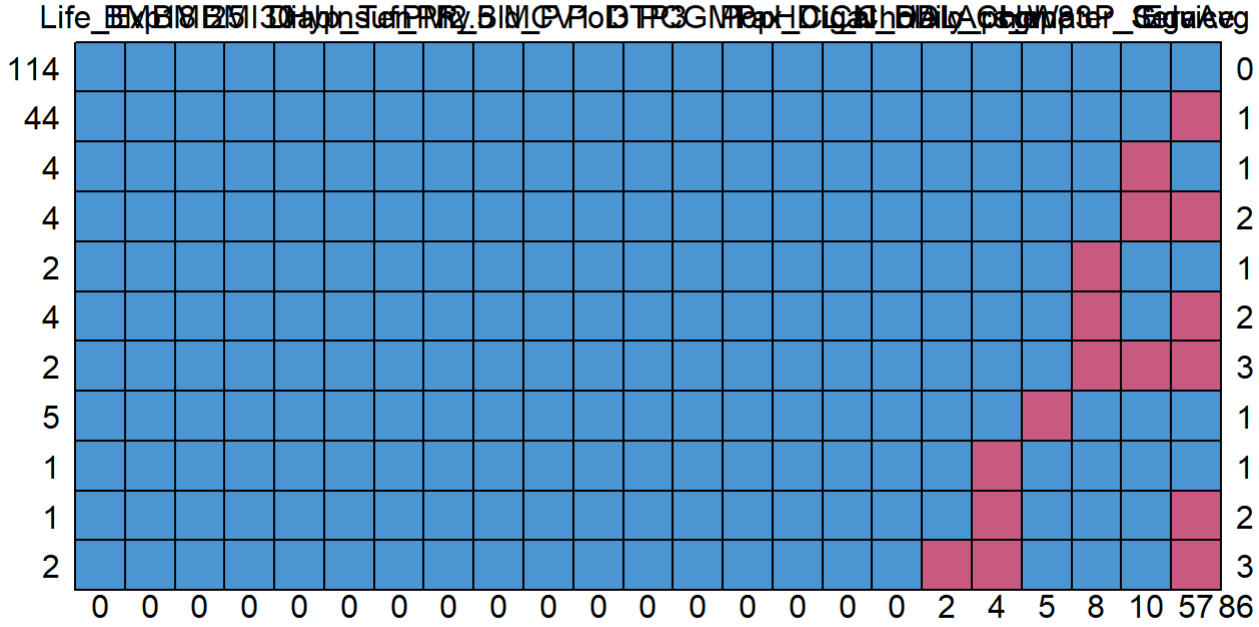
```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_lasso would be chosen as imputation.

```
#replace with imputed data: Tax_Ciga

Re_imp_df184$Tax_Ciga <- mice_imputed$imputed_lasso
md.pattern(Re_imp_df184)
```


114
44
4
4
2
4
2
5
1
1
2
0
1
1
2
1
2
3
1
1
2
3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 4 5 8 10 57 86

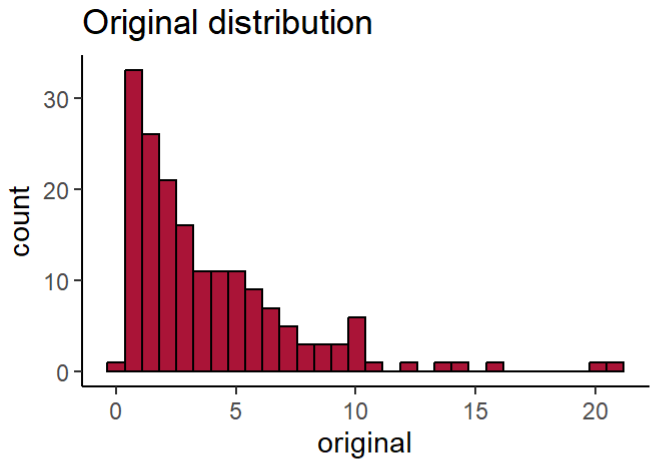
```
##     Life_Exp BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy PM2.5 R_Bld_P MCV1 Pol3
## 114        1     1     1     1    1       1         1     1       1    1    1
## 44         1     1     1     1    1       1         1     1       1    1    1
## 4          1     1     1     1    1       1         1     1       1    1    1
## 4          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 4          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
## 5          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 1          1     1     1     1    1       1         1     1       1    1    1
## 2          1     1     1     1    1       1         1     1       1    1    1
##            0     0     0     0    0       0         0     0       0    0    0
##     DTP3 PCGMP Pop Tax_Ciga HDL_Chol Cal_Daily N_HDL_Chol Alc_pcgmp Hep83
## 114    1     1   1        1        1         1          1         1     1
## 44     1     1   1        1        1         1          1         1     1
## 4      1     1   1        1        1         1          1         1     1
## 4      1     1   1        1        1         1          1         1     1
## 2      1     1   1        1        1         1          1         1     1
## 4      1     1   1        1        1         1          1         1     1
## 2      1     1   1        1        1         1          1         1     1
## 5      1     1   1        1        1         1          1         1     0
## 1      1     1   1        1        1         1          1         0     1
## 1      1     1   1        1        1         1          1         0     1
## 2      1     1   1        1        1         1          0         0     1
##        0     0   0        0        0         0          2         4     5
##     Acs_Water_Service P_Ciga EduAvg
## 114                 1      1      1  0
## 44                  1      1      0  1
## 4                   1      0      1  1
## 4                   1      0      0  2
## 2                   0      1      1  1
## 4                   0      1      0  2
## 2                   0      0      0  3
## 5                   1      1      1  1
## 1                   1      1      1  1
## 1                   1      1      0  2
## 2                   1      1      0  3
##                     8     10     57 86
```

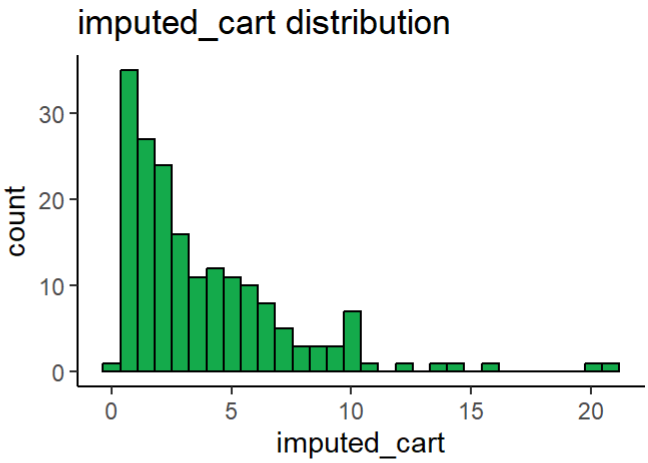
6) imputation on P_Ciga

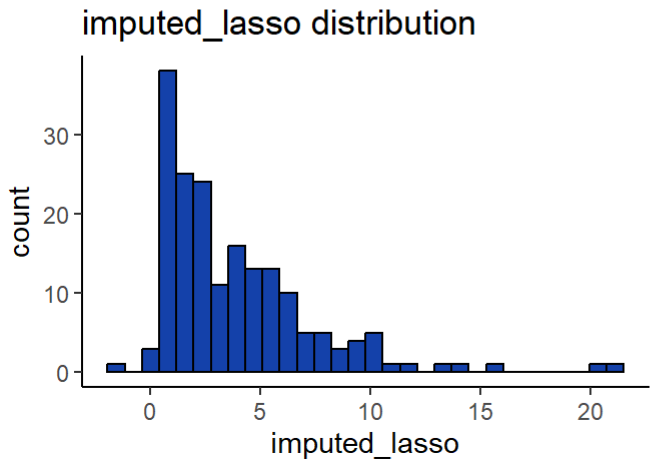
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```







From above, imputed_cart would be chosen as imputation.

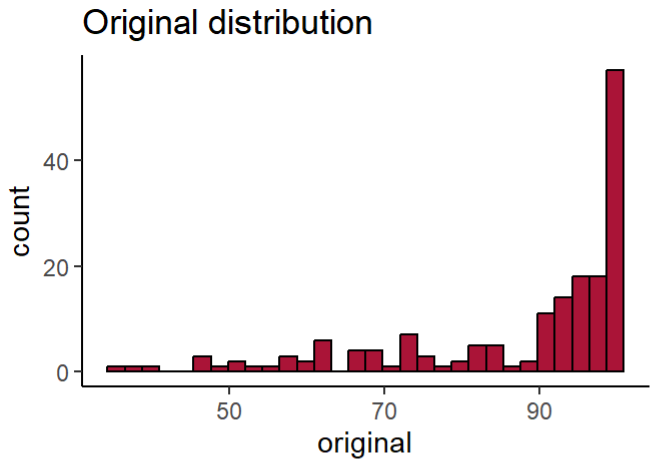
```
#replace with imputed data: P_Ciga

Re_imp_df184$P_Ciga <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

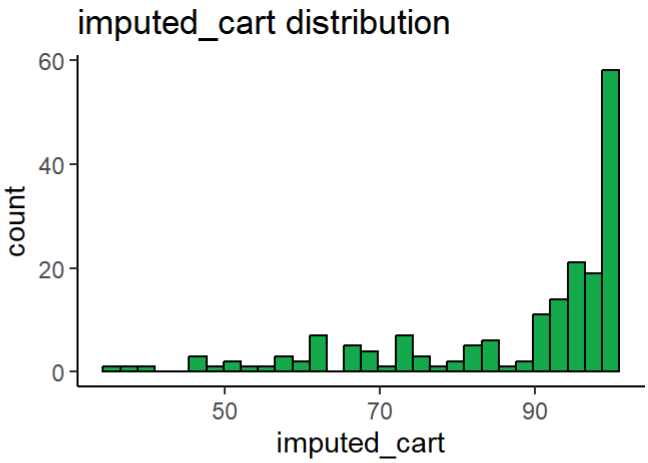
7) imputation on Acs_Water_Service

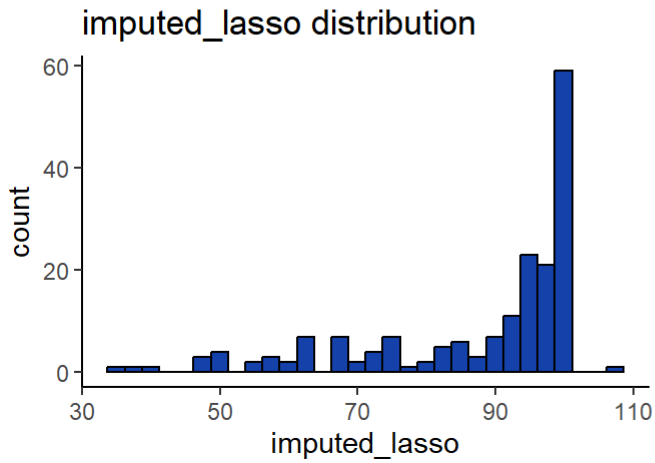
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

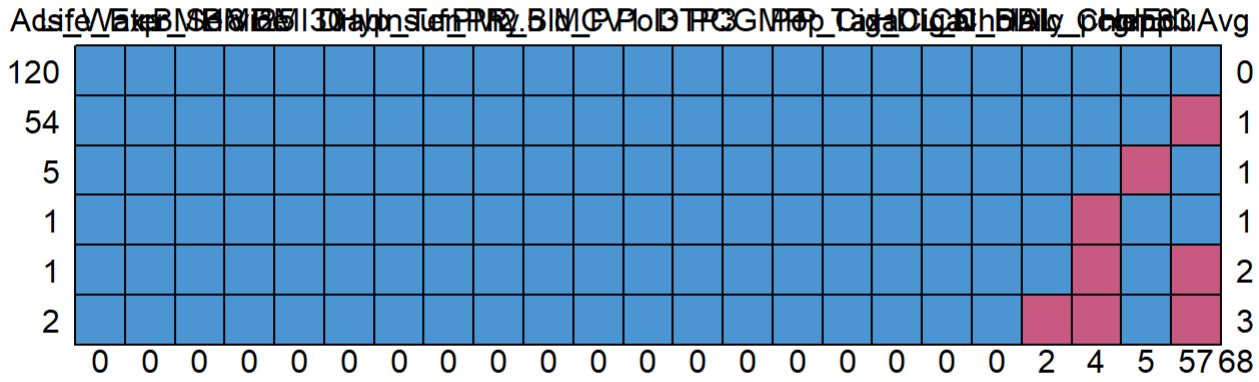






From above, imputed_cart would be chosen as imputation.

```
#replace with imputed data: Acs_Water_Service

Re_imp_df184$Acs_Water_Service <- mice_imputed$imputed_cart
md.pattern(Re_imp_df184)
```



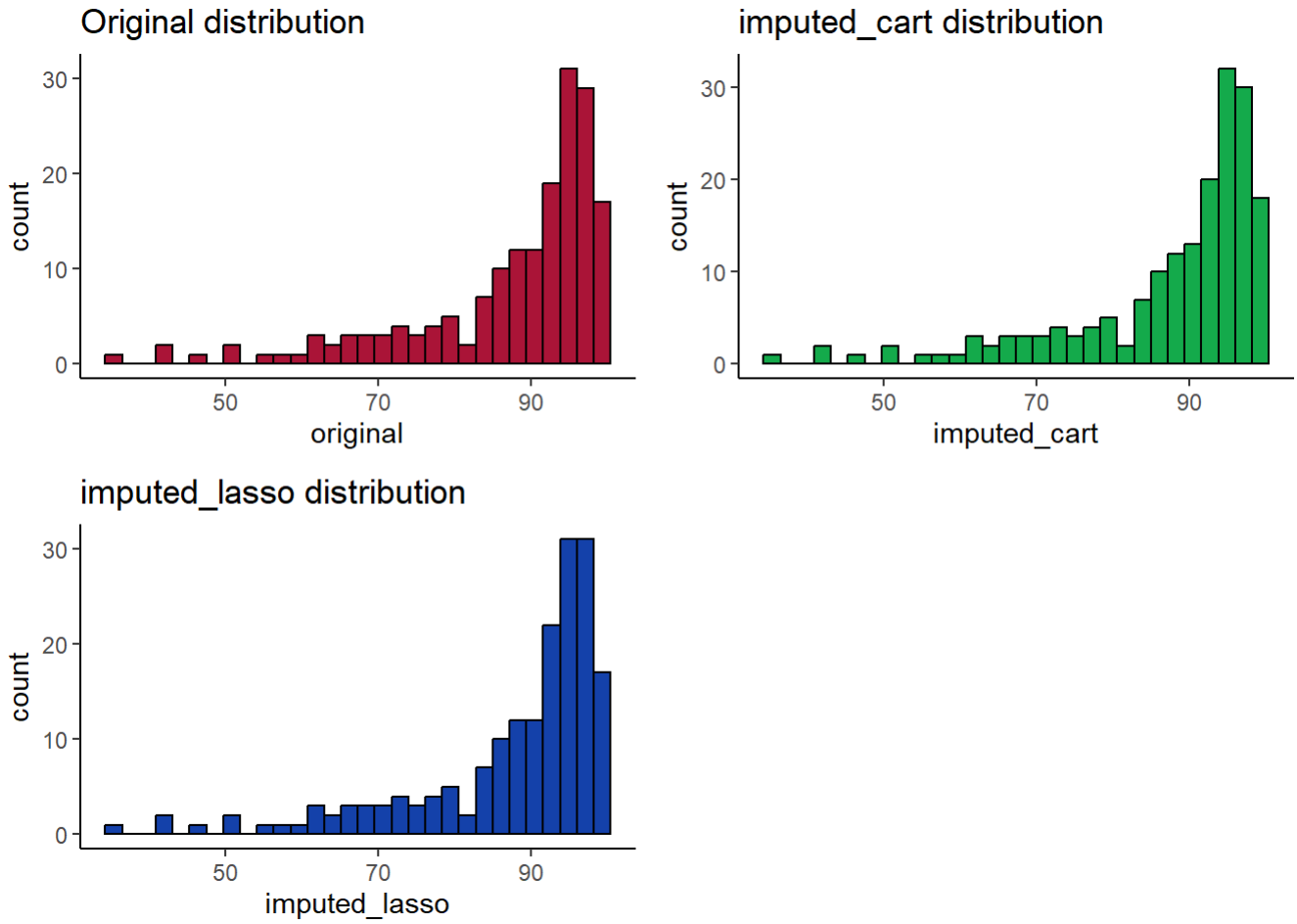

```
##     Life_Exp Acs_Water_Service BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy PM2.5
## 120        1                 1     1     1     1    1       1         1     1
## 54         1                 1     1     1     1    1       1         1     1
## 5          1                 1     1     1     1    1       1         1     1
## 1          1                 1     1     1     1    1       1         1     1
## 1          1                 1     1     1     1    1       1         1     1
## 2          1                 1     1     1     1    1       1         1     1
##            0                 0     0     0     0    0       0         0     0
##     R_Bld_P MCV1 Pol3 DTP3 PCGMP Pop P_Ciga Tax_Ciga HDL_Chol Cal_Daily
## 120       1    1    1    1     1   1      1        1        1         1
## 54        1    1    1    1     1   1      1        1        1         1
## 5         1    1    1    1     1   1      1        1        1         1
## 1         1    1    1    1     1   1      1        1        1         1
## 1         1    1    1    1     1   1      1        1        1         1
## 2         1    1    1    1     1   1      1        1        1         1
##           0    0    0    0     0   0      0        0        0         0
##     N_HDL_Chol Alc_pcgmp Hep83 EduAvg
## 120          1         1     1      1  0
## 54           1         1     1      0  1
## 5            1         1     0      1  1
## 1            1         0     1      1  1
## 1            1         0     1      0  2
## 2            0         0     1      0  3
##              2         4     5     57 68
```

8) imputation on Hep83

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

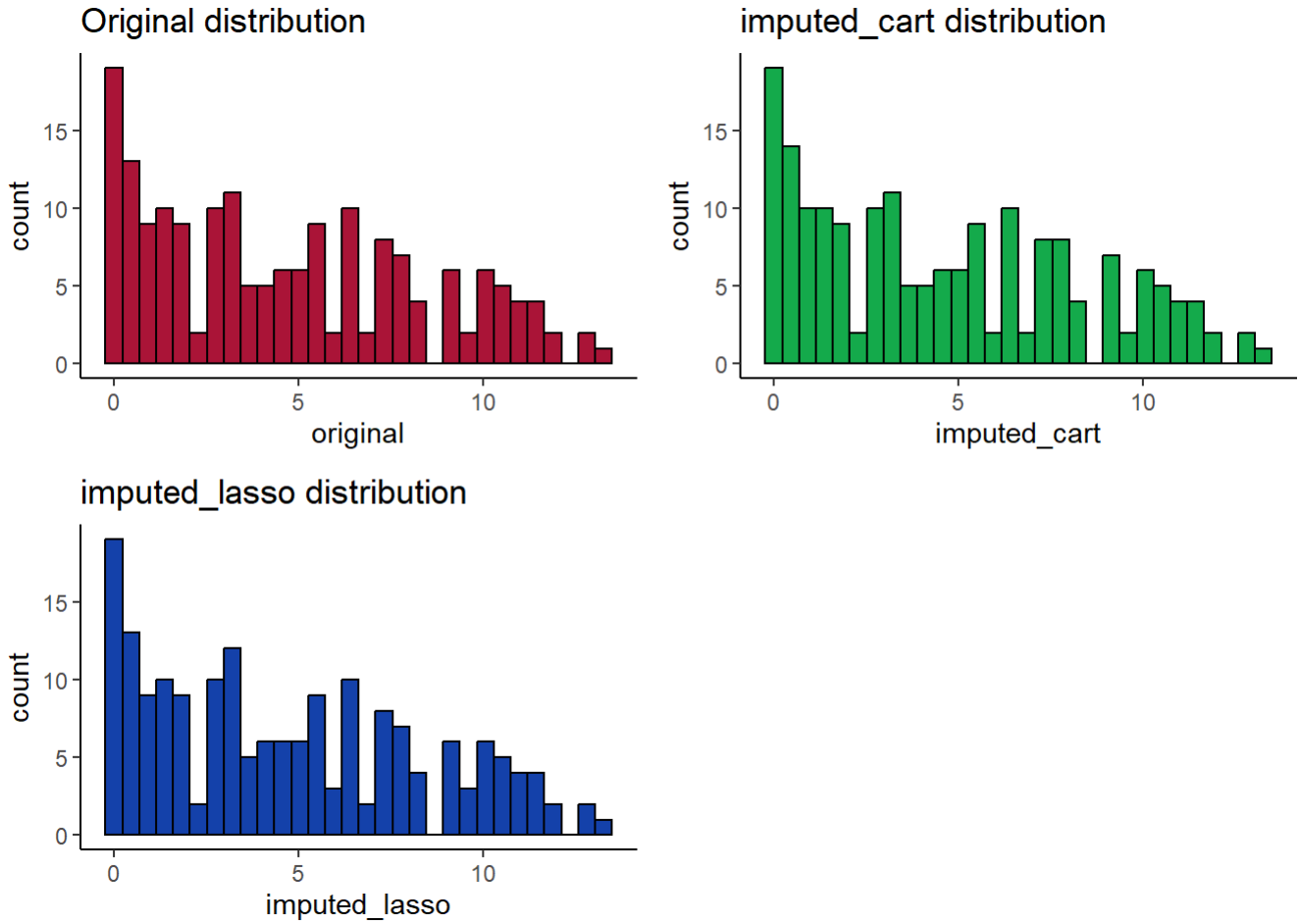
```
#replace with imputed data: Hep83

Re_imp_df184$Hep83 <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

9) imputation on Alc_pcgmp

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

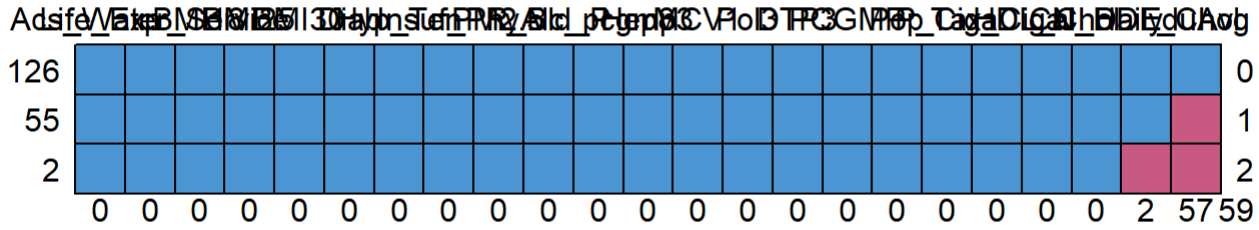
```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_lasso would be chosen as imputation.

```
#replace with imputed data: Alc_pcgmp

Re_imp_df184$Alc_pcgmp <- mice_imputed$imputed_lasso
md.pattern(Re_imp_df184)
```



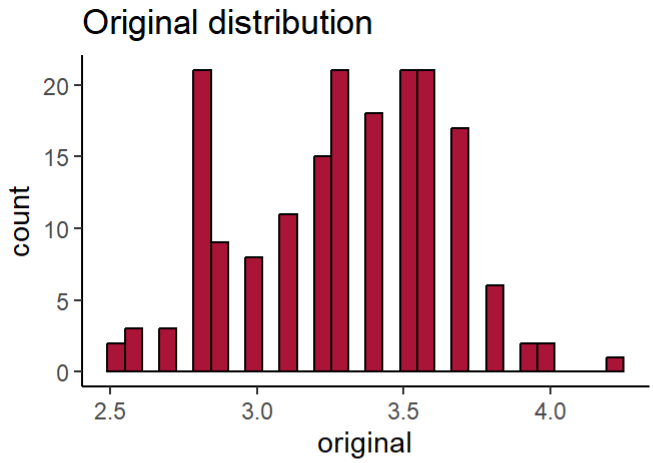

```
##     Life_Exp Acs_Water_Service BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy PM2.5
## 126        1                 1     1     1     1    1       1         1     1
## 55         1                 1     1     1     1    1       1         1     1
## 2          1                 1     1     1     1    1       1         1     1
##            0                 0     0     0     0    0       0         0     0
##     R_Bld_P Alc_pcgmp Hep83 MCV1 Pol3 DTP3 PCGMP Pop P_Ciga Tax_Ciga HDL_Chol
## 126       1         1     1    1    1    1     1   1      1        1        1
## 55        1         1     1    1    1    1     1   1      1        1        1
## 2         1         1     1    1    1    1     1   1      1        1        1
##           0         0     0    0    0    0     0   0      0        0        0
##     Cal_Daily N_HDL_Chol EduAvg
## 126         1          1      1  0
## 55          1          1      0  1
## 2           1          0      0  2
##             0          2     57 59
```

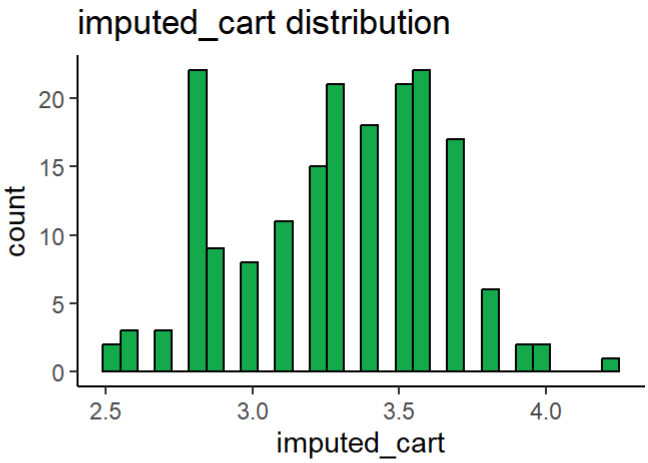
10) imputation on N_HDL_Chol

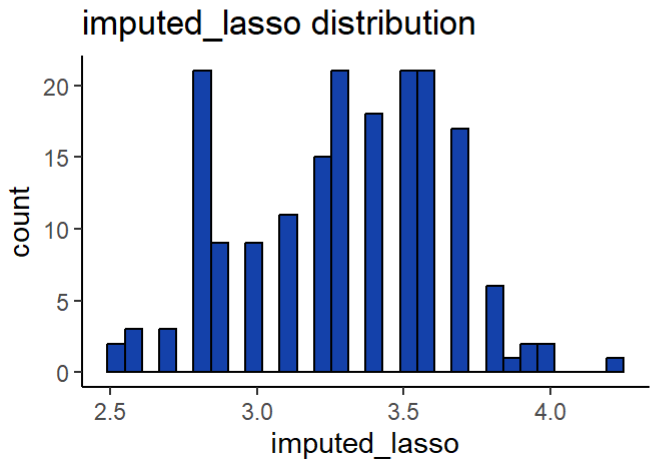
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

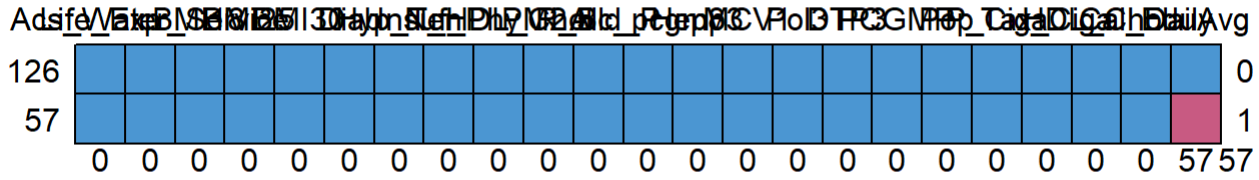






From above, imputed_cart would be chosen as imputation.

```
#replace with imputed data: N_HDL_Chol

Re_imp_df184$N_HDL_Chol <- mice_imputed$imputed_lasso
md.pattern(Re_imp_df184)
```



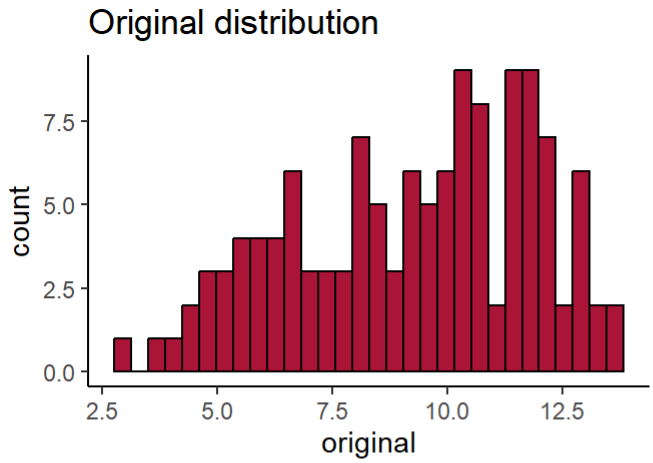

```
##     Life_Exp Acs_Water_Service BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy
## 126        1                 1     1     1     1    1       1         1
## 57         1                 1     1     1     1    1       1         1
##            0                 0     0     0     0    0       0         0
##     N_HDL_Chol PM2.5 R_Bld_P Alc_pcgmp Hep83 MCV1 Pol3 DTP3 PCGMP Pop P_Ciga
## 126          1     1       1         1     1    1    1    1     1   1      1
## 57           1     1       1         1     1    1    1    1     1   1      1
##              0     0       0         0     0    0    0    0     0   0      0
##     Tax_Ciga HDL_Chol Cal_Daily EduAvg
## 126        1        1         1      1  0
## 57         1        1         1      0  1
##            0        0         0     57 57
```

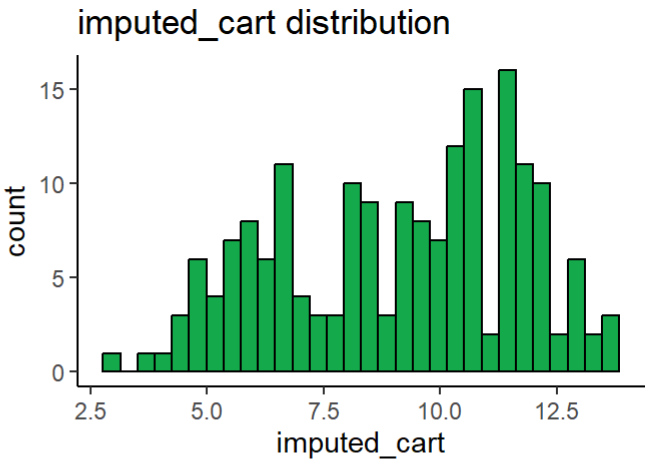
11) imputation on EduAvg

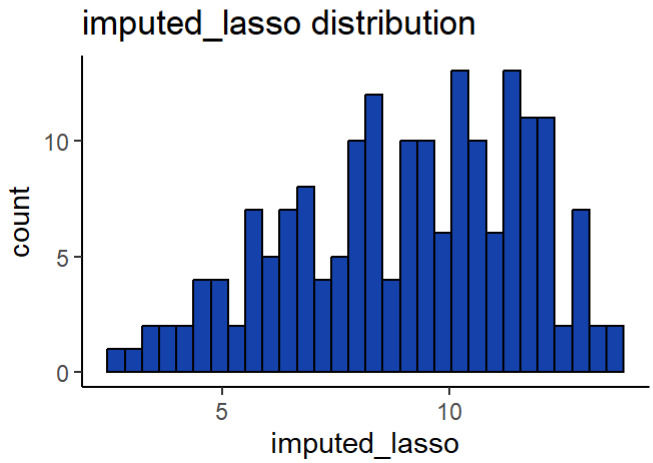
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 57 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```







From above, imputed_cart would be chosen as imputation.

```
#replace with imputed data: EduAvg

Re_imp_df184$EduAvg <- mice_imputed$imputed_cart
md.pattern(Re_imp_df184)
```

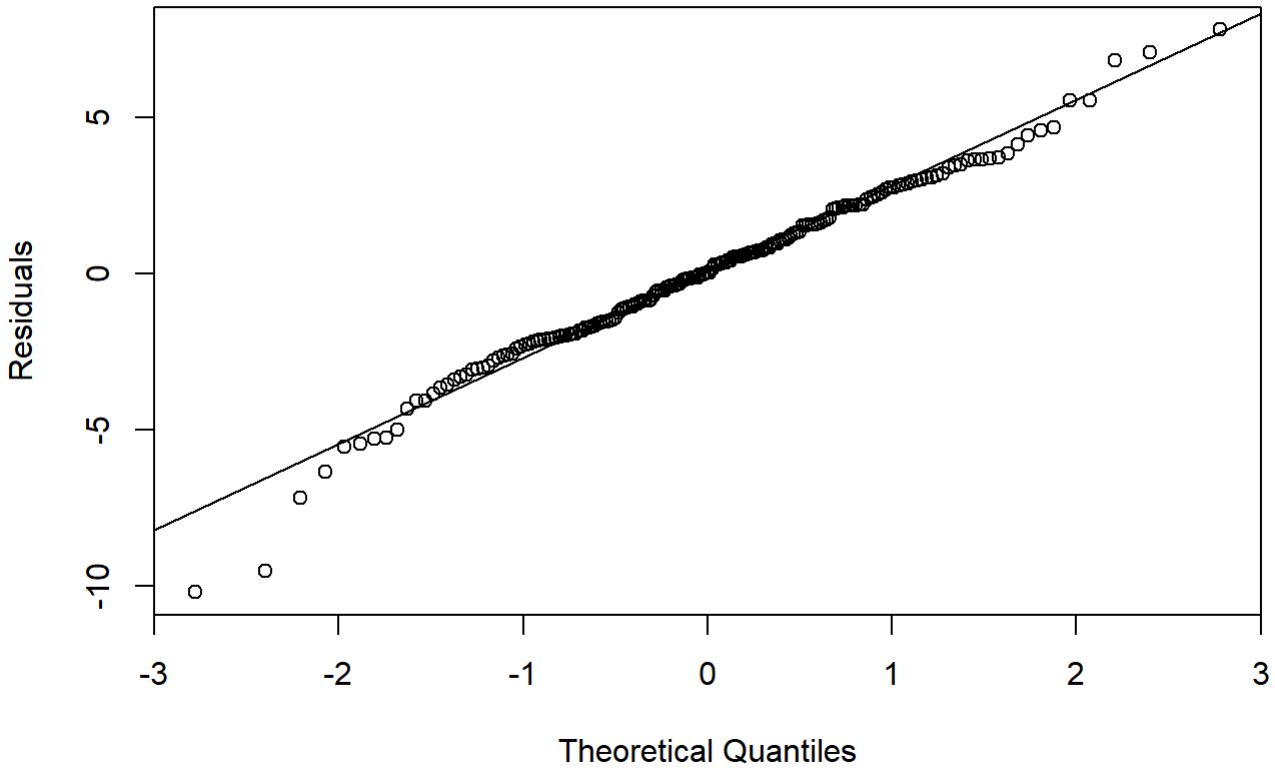
```
##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```

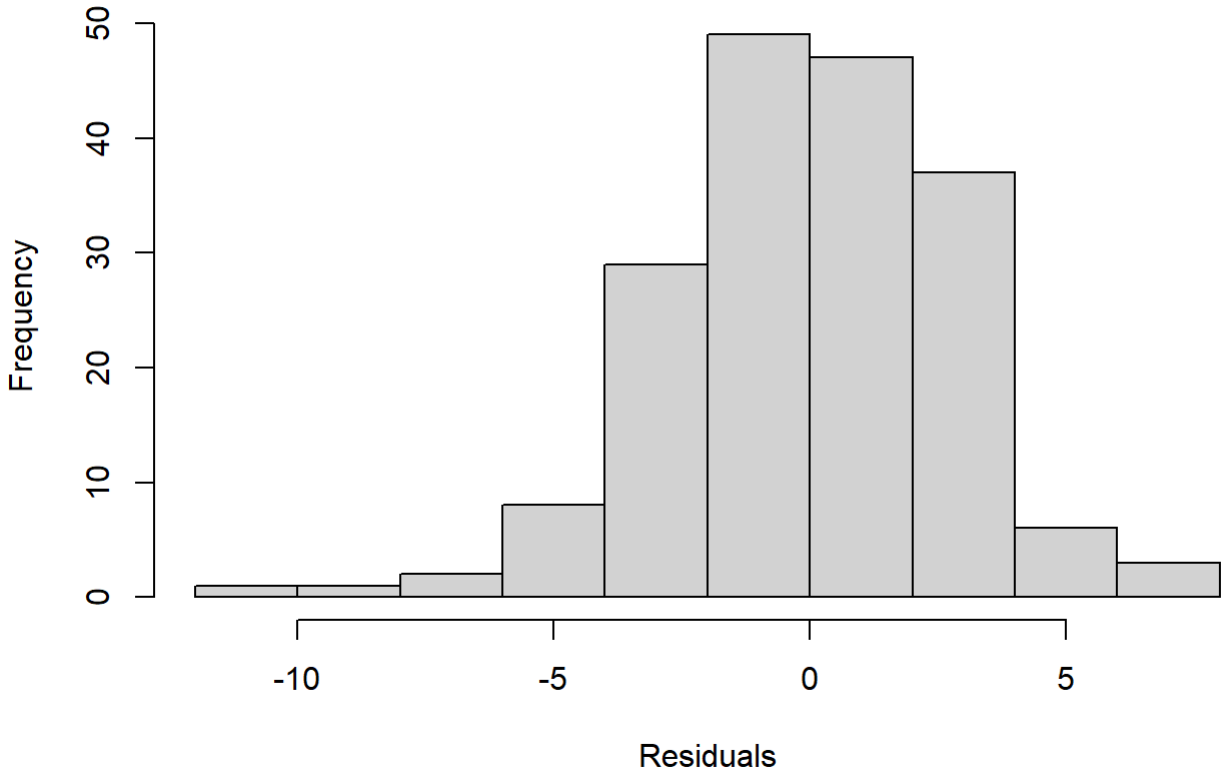

```
##     Life_Exp Acs_Water_Service BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy
## 183        1                 1     1     1     1    1       1         1
##            0                 0     0     0     0    0       0         0
##     N_HDL_Chol PM2.5 R_Bld_P Alc_pcgmp Hep83 MCV1 Pol3 DTP3 PCGMP Pop EduAvg
## 183          1     1       1         1     1    1    1    1     1   1      1
##              0     0       0         0     0    0    0    0     0   0      0
##     P_Ciga Tax_Ciga HDL_Chol Cal_Daily
## 183      1        1        1         1 0
##          0        0        0         0 0
```

Model Summary after completing imputation regression

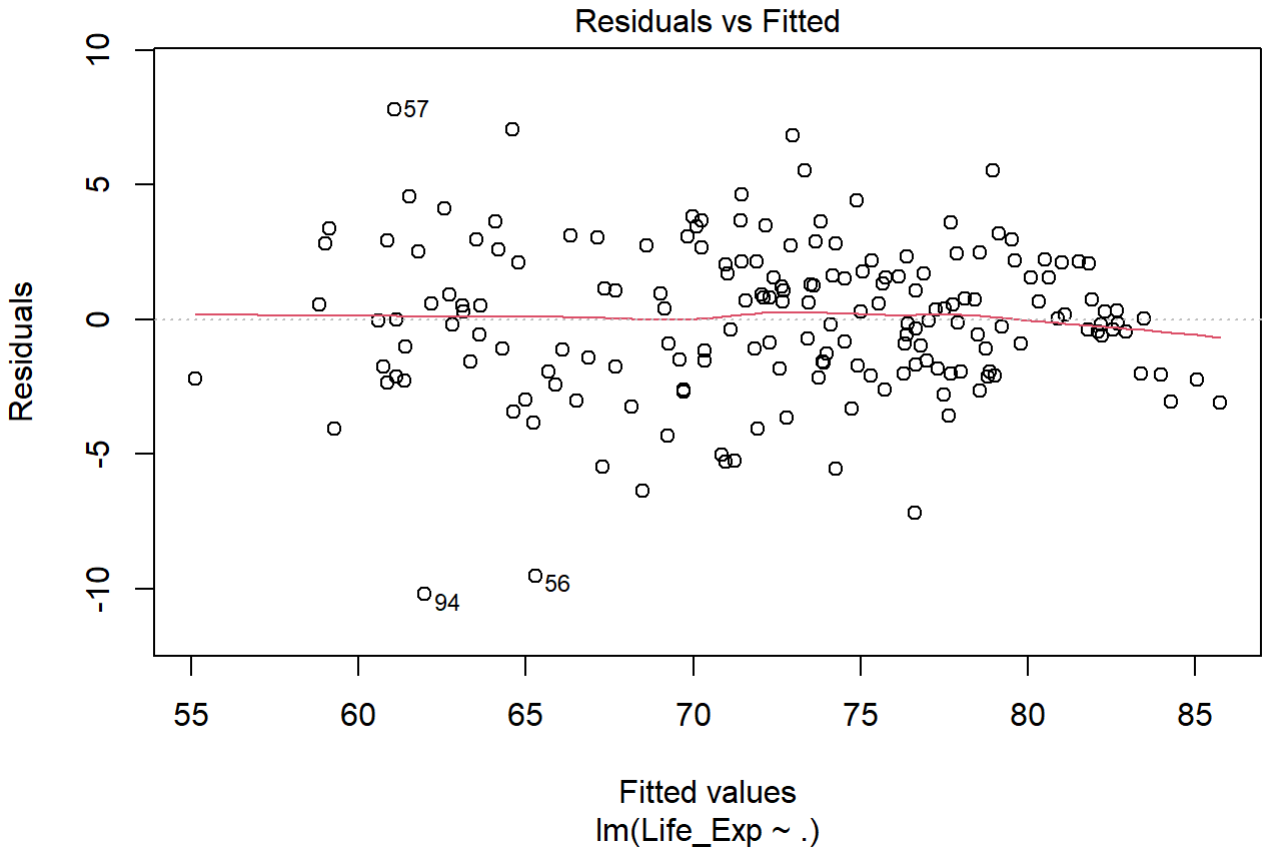
```
##
## Call:
## lm(formula = Life_Exp ~ ., data = imp_df_184)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2017  -1.7927   0.0147   1.9200   7.7933
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.359e+01  4.289e+00   7.831 6.43e-13 ***
## X                 -1.508e-03  4.490e-03  -0.336 0.737435
## Acs_Water_Service  1.182e-01  3.016e-02   3.920 0.000131 ***
## BMI18              5.112e-02  1.014e-01   0.504 0.615013
## BMI25              1.804e-01  7.168e-02   2.516 0.012855 *
## BMI30             -2.278e-01  7.727e-02  -2.948 0.003681 **
## Diab              -1.076e-01  6.259e-02  -1.719 0.087593 .
## Hyp_Ten           -1.536e-01  1.106e-01  -1.390 0.166594
## Insuf_Phy          1.023e-01  2.893e-02   3.537 0.000531 ***
## N_HDL_Chol         3.375e+00  9.699e-01   3.480 0.000647 ***
## PM2.5              6.158e-03  2.342e-02   0.263 0.792897
## R_Bld_P            1.277e-02  1.016e-01   0.126 0.900165
## Alc_pcgmp          3.506e-02  1.005e-01   0.349 0.727631
## Hep83              1.975e-02  7.345e-02   0.269 0.788393
## MCV1               1.453e-02  3.742e-02   0.388 0.698344
## Pol3               9.764e-02  9.445e-02   1.034 0.302802
## DTP3              -6.288e-02  1.113e-01  -0.565 0.572989
## PCGMP              6.590e-05  2.343e-05   2.813 0.005528 **
## Pop               -9.224e-10  1.655e-09  -0.557 0.578142
## EduAvg            -3.546e-02  1.763e-01  -0.201 0.840790
## P_Ciga             2.640e-02  1.124e-01   0.235 0.814619
## Tax_Ciga           3.916e+00  1.252e+00   3.129 0.002089 **
## HDL_Chol           7.877e-01  2.635e+00   0.299 0.765353
## Cal_Daily          2.425e-03  8.406e-04   2.885 0.004454 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.971 on 159 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8337
## F-statistic: 40.66 on 23 and 159 DF,  p-value: < 2.2e-16
```

```
## AIC test:  942.2016
```

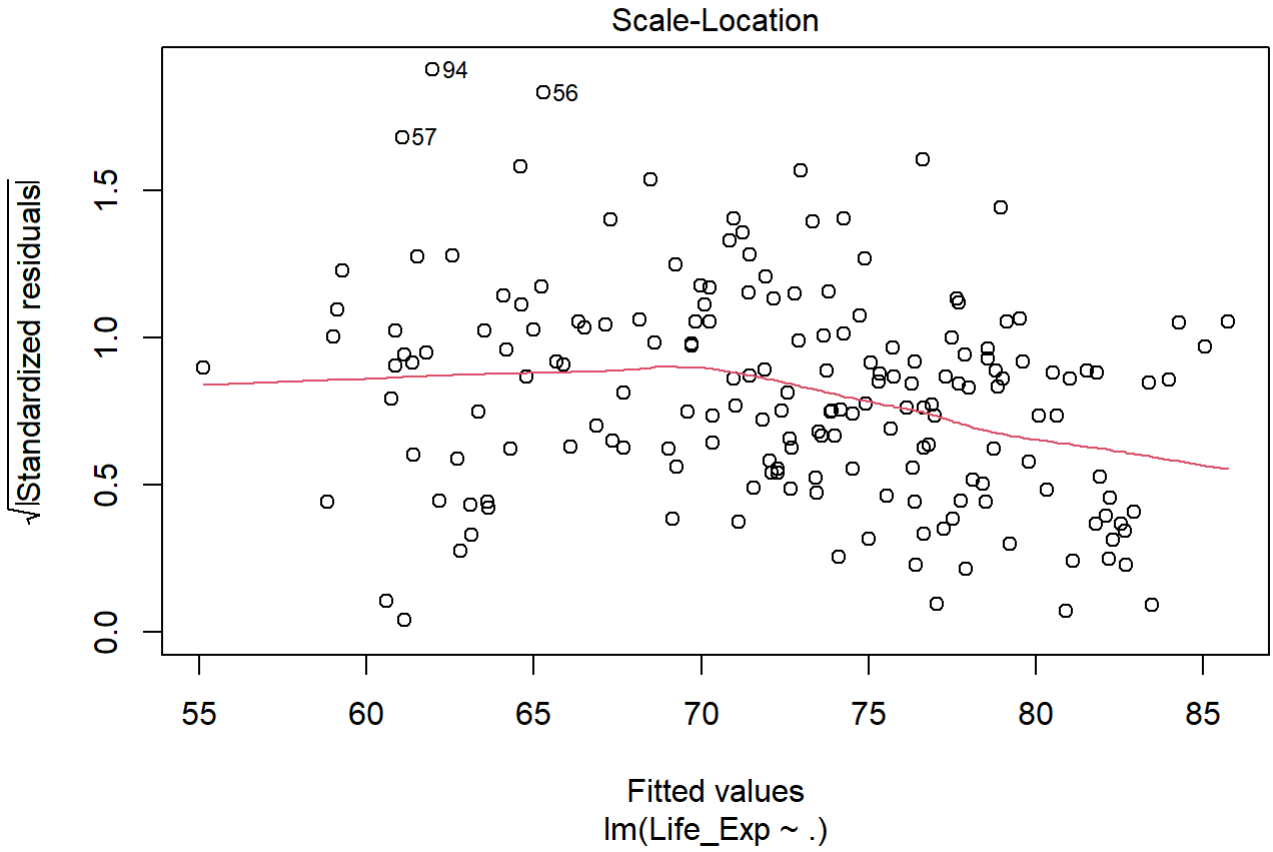

Residuals
5
0
-5
-10
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Frequency
50
40
30
20
10
0
-10
-5
0
5
Residuals

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.98327, p-value = 0.02755
```




```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 2.0902, p-value = 0.7005
## alternative hypothesis: true autocorrelation is greater than 0
```



Backward Elimination on Approach 3

In each round of backward elimination, we chose the predictor with highest p-value (i.e. least significance on model) and eliminate them. AIC score of model after elimination will be calculated for determining when to hault the elimination.

R1 - remove highest: R_Bld_P

```
## AIC test:  940.2198
```

R2 - remove highest:EduAvg

```
## AIC test:  938.2682
```

R3 - remove highest: P_Ciga

```
## AIC test:  936.3316
```

R4 - remove highest: HDL_Chol

```
## AIC test:  934.4028
```

R5 - remove highest: PM2.5

```
## AIC test:  932.4947
```

R6 - remove highest: MCV1

```
## AIC test:  930.5932
```

R7 - remove highest: Alc_pcgmp

```
## AIC test:  928.6987
```

R8 - remove highest: BMI18

```
## AIC test:  926.8999
```

R9 - remove highest: Hep83

```
## AIC test:  925.172
```

R10 - remove highest: DTP3

```
## AIC test:  923.3664
```

R11 - remove highest: Pop

```
## AIC test:  921.6088
```

```
##
## Call:
## lm(formula = Life_Exp ~ Acs_Water_Service + BMI25 + BMI30 + Diab +
##     Hyp_Ten + Insuf_Phy + N_HDL_Chol + Pol3 + PCGMP + Tax_Ciga +
##     Cal_Daily, data = imp_df_184)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2613  -1.7037   0.0616   1.8719   7.9963
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.514e+01  2.967e+00  11.842  < 2e-16 ***
## Acs_Water_Service  1.151e-01  2.647e-02   4.346 2.37e-05 ***
## BMI25              1.607e-01  5.140e-02   3.127 0.002077 **
## BMI30             -2.157e-01  6.650e-02  -3.244 0.001418 **
## Diab              -1.059e-01  5.097e-02  -2.078 0.039238 *
## Hyp_Ten           -1.311e-01  3.722e-02  -3.523 0.000548 ***
## Insuf_Phy          1.014e-01  2.267e-02   4.472 1.41e-05 ***
## N_HDL_Chol         3.366e+00  8.354e-01   4.029 8.41e-05 ***
## Pol3               6.817e-02  2.139e-02   3.187 0.001708 **
## PCGMP              7.224e-05  1.564e-05   4.618 7.60e-06 ***
## Tax_Ciga           3.888e+00  1.163e+00   3.343 0.001020 **
## Cal_Daily          2.482e-03  7.964e-04   3.116 0.002148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.878 on 171 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8439
## F-statistic: 90.46 on 11 and 171 DF,  p-value: < 2.2e-16
```

```
## AIC test:  919.8766
```

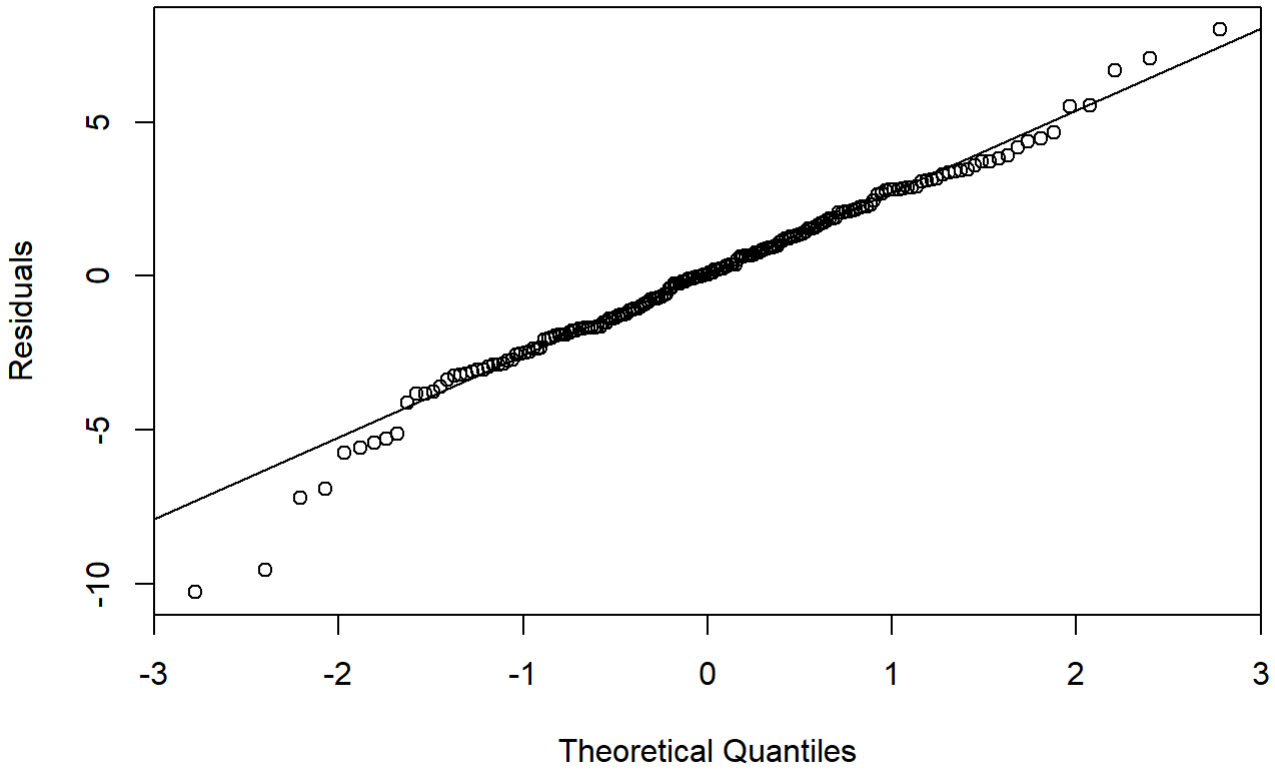
Observation:

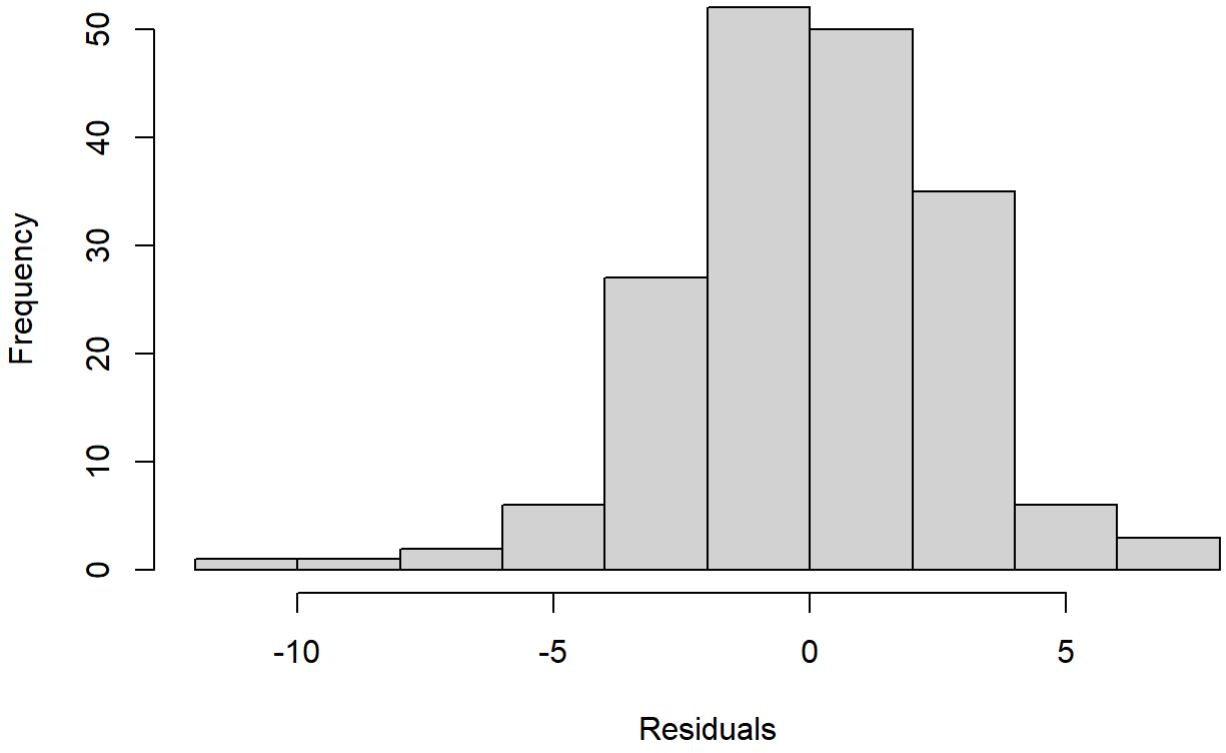
As we can see, all predictors remaining have p-value <0.05 in F-test, thus we can halt back elimination.

Thus, the final model is as follows:

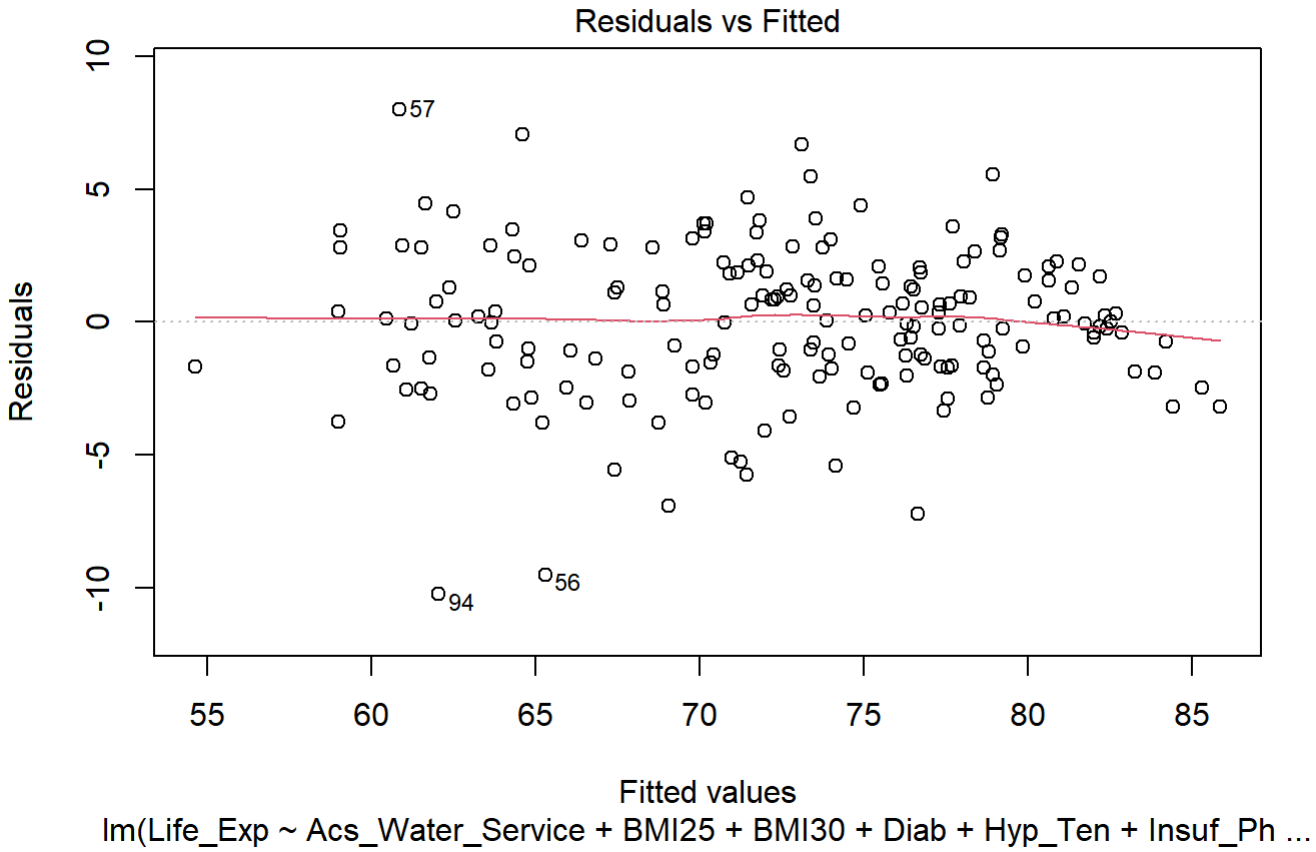
```
##
## Call:
## lm(formula = Life_Exp ~ Acs_Water_Service + BMI25 + BMI30 + Diab +
##     Hyp_Ten + Insuf_Phy + N_HDL_Chol + Pol3 + PCGMP + Tax_Ciga +
##     Cal_Daily, data = imp_df_184)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2613  -1.7037   0.0616   1.8719   7.9963
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.514e+01  2.967e+00  11.842  < 2e-16 ***
## Acs_Water_Service  1.151e-01  2.647e-02   4.346 2.37e-05 ***
## BMI25              1.607e-01  5.140e-02   3.127 0.002077 **
## BMI30             -2.157e-01  6.650e-02  -3.244 0.001418 **
## Diab              -1.059e-01  5.097e-02  -2.078 0.039238 *
## Hyp_Ten           -1.311e-01  3.722e-02  -3.523 0.000548 ***
## Insuf_Phy          1.014e-01  2.267e-02   4.472 1.41e-05 ***
## N_HDL_Chol         3.366e+00  8.354e-01   4.029 8.41e-05 ***
## Pol3               6.817e-02  2.139e-02   3.187 0.001708 **
## PCGMP              7.224e-05  1.564e-05   4.618 7.60e-06 ***
## Tax_Ciga           3.888e+00  1.163e+00   3.343 0.001020 **
## Cal_Daily          2.482e-03  7.964e-04   3.116 0.002148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.878 on 171 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8439
## F-statistic: 90.46 on 11 and 171 DF,  p-value: < 2.2e-16
```

```
## AIC test:  919.8766
```

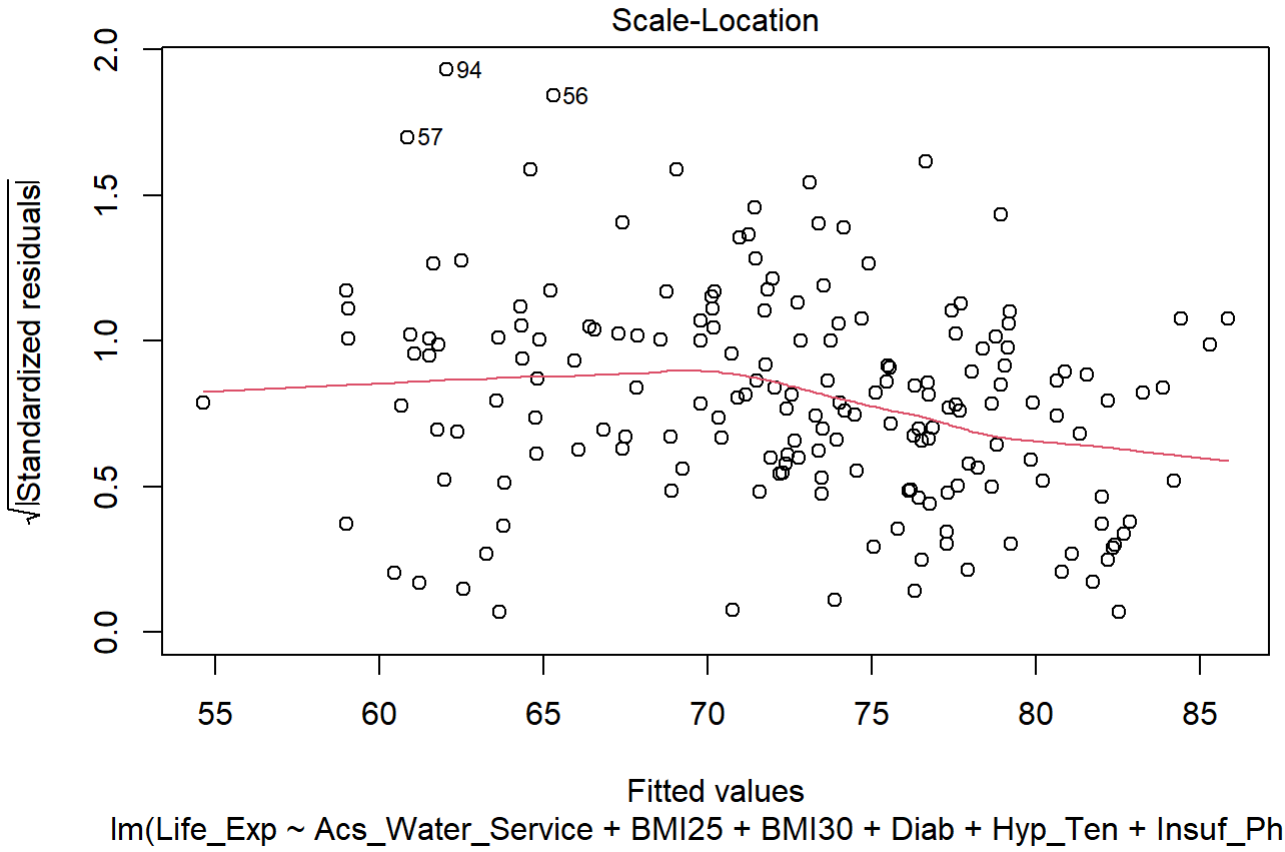

Residuals
5
0
-5
-10
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Frequency
50
40
30
20
10
0
-10
-5
0
5
Residuals

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.98235, p-value = 0.02063
```




```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 2.0894, p-value = 0.7303
## alternative hypothesis: true autocorrelation is greater than 0
```



Model Comparison Metrics Summary

```
library(flexmix)
```

```
## Warning: package 'flexmix' was built under R version 4.4.2
```

```
## Loading required package: lattice
```

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.4.2
```

```
significant_C <- tidy(BE_com_lmod) %>%
  filter(p.value < 0.05) %>%
  select(term, p.value)

cat("Model Comparison Result of Models:", "\n\n")
```

```
## Model Comparison Result of Models:
```

```
cat("1) Model with Only Complete Case","\n\n")
```

```
## 1) Model with Only Complete Case
```

```
cat("   Adjusted R squared: ", glance(BE_com_lmod)$adj.r.squared , "\n")
```

```
##    Adjusted R squared:  0.8368908
```

```
cat("   AIC score:          ",AIC(BE_com_lmod),"\n")
```

```
##    AIC score:           548.9232
```

```
cat("   BIC score:          ",BIC(BE_com_lmod),"\n\n")
```

```
##    BIC score:           594.2017
```

```
cat("   Predictor with p-value< 0.05:","\n")
```

```
##    Predictor with p-value< 0.05:
```

```
print(significant_C)
```

```
## # A tibble: 8 × 2
##   term              p.value
##   <chr>               <dbl>
## 1 (Intercept) 0.00000000200
## 2 BMI25       0.0199
## 3 Diab        0.0257
## 4 Hyp_Ten     0.00112
## 5 Insuf_Phy   0.00737
## 6 N_HDL_Chol  0.00284
## 7 PCGMP       0.00305
## 8 Cal_Daily   0.00595
```

```
library(flexmix)
library(broom)

significant_M <- tidy(BE_M_lmod) %>%
  filter(p.value < 0.05) %>%
  select(term, p.value)

cat("2) Model imputed with mean","\n\n")
```

```
## 2) Model imputed with mean
```

```
cat("   Adjusted R squared: ", glance(BE_M_lmod)$adj.r.squared , "\n")
```

```
##    Adjusted R squared:  0.8291135
```

```
cat("   AIC score:          ",AIC(BE_M_lmod),"\n")
```

```
##    AIC score:           936.4524
```

```
cat("   BIC score:          ",BIC(BE_M_lmod),"\n\n")
```

```
##    BIC score:           978.1757
```

```
cat("   Predictor with p-value< 0.05:","\n")
```

```
##    Predictor with p-value< 0.05:
```

```
print(significant_M)
```

```
## # A tibble: 11 × 2
##    term               p.value
##    <chr>                <dbl>
##  1 (Intercept)       3.66e-22
##  2 Acs_Water_Service 8.03e- 6
##  3 BMI25             1.43e- 4
##  4 BMI30             2.87e- 4
##  5 Diab              7.73e- 3
##  6 Hyp_Ten           1.22e- 3
##  7 Insuf_Phy         1.32e- 5
##  8 N_HDL_Chol        4.30e- 4
##  9 Pol3              9.15e- 4
## 10 PCGMP             1.20e- 5
## 11 Tax_Ciga          1.63e- 3
```

```
library(flexmix)
library(broom)

significant_R <- tidy(BE_imp_lmod) %>%
  filter(p.value < 0.05) %>%
  select(term, p.value)

cat("3) Model imputed with Regression","\n\n")
```

```
## 3) Model imputed with Regression
```

```
cat("   Adjusted R squared: ", glance(BE_imp_lmod)$adj.r.squared , "\n")
```

```
##    Adjusted R squared:  0.8439117
```

```
cat("   AIC score:          ",AIC(BE_imp_lmod),"\n")
```

```
##    AIC score:          919.8766
```

```
cat("   BIC score:          ",BIC(BE_imp_lmod),"\n\n")
```

```
##    BIC score:          961.5999
```

```
cat("   Predictor with p-value< 0.05:","\n")
```

```
##    Predictor with p-value< 0.05:
```

```
print(significant_R)
```

```
## # A tibble: 12 × 2
##    term               p.value
##    <chr>                <dbl>
##  1 (Intercept)       5.23e-24
##  2 Acs_Water_Service 2.37e- 5
##  3 BMI25             2.08e- 3
##  4 BMI30             1.42e- 3
##  5 Diab              3.92e- 2
##  6 Hyp_Ten           5.48e- 4
##  7 Insuf_Phy         1.41e- 5
##  8 N_HDL_Chol        8.41e- 5
##  9 Pol3              1.71e- 3
## 10 PCGMP             7.60e- 6
## 11 Tax_Ciga          1.02e- 3
## 12 Cal_Daily         2.15e- 3
```

Transformation of A3 model

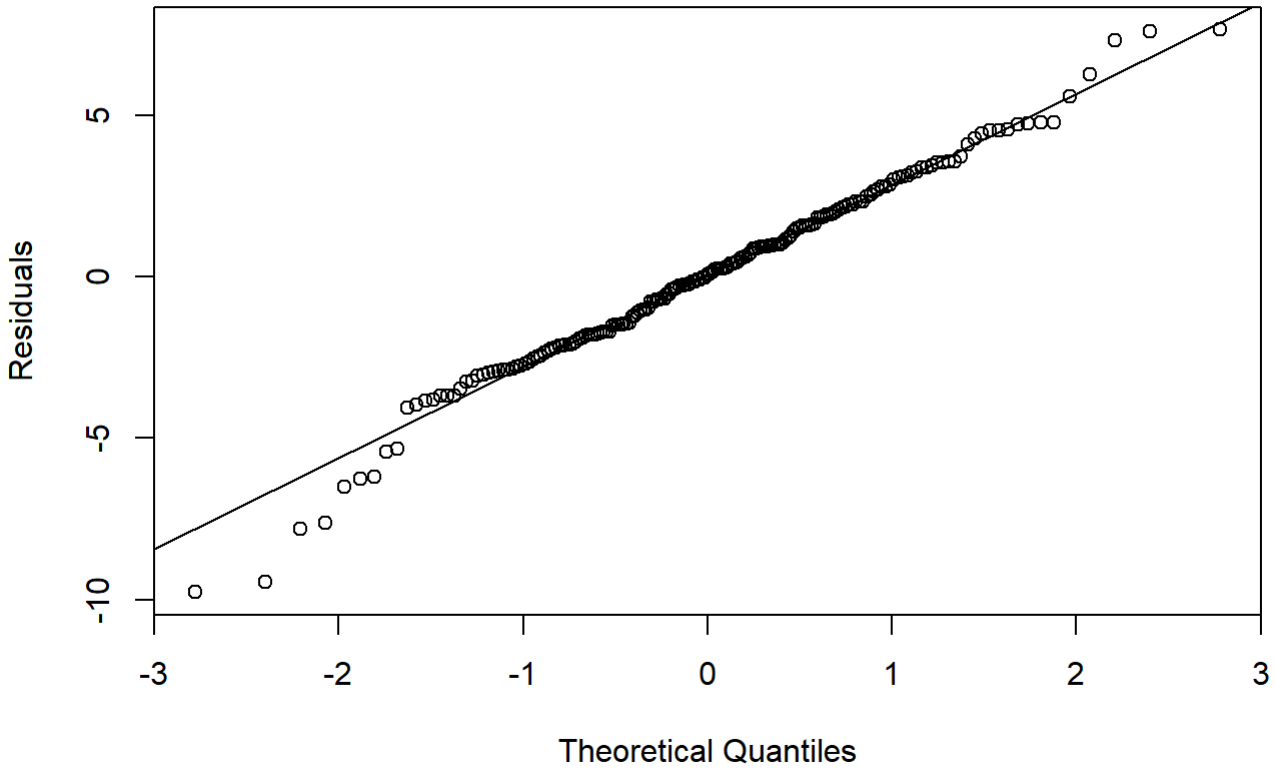
```
recip_PCGMP <- 1/imp_df_184$PCGMP
sqrt_Cal_Daily <- sqrt(imp_df_184$Cal_Daily)

tran_BE_Imp_lmod <- lm(Life_Exp ~ (Acs_Water_Service) + BMI25 + BMI30 + Diab +
    (Hyp_Ten) + (Insuf_Phy) + (N_HDL_Chol) + (Pol3) + recip_PCGMP + (Tax_Ciga) + sqrt_Cal_Dai
ly, data = imp_df_184)

test_mod <- tran_BE_Imp_lmod

summary(test_mod)
```

```
##
## Call:
## lm(formula = Life_Exp ~ (Acs_Water_Service) + BMI25 + BMI30 +
##     Diab + (Hyp_Ten) + (Insuf_Phy) + (N_HDL_Chol) + (Pol3) +
##     recip_PCGMP + (Tax_Ciga) + sqrt_Cal_Daily, data = imp_df_184)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7710 -1.8647  0.0521  1.9361  7.6427
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        23.68386    5.18443   4.568 9.38e-06 ***
## Acs_Water_Service   0.12496    0.02913   4.290 2.98e-05 ***
## BMI25               0.17711    0.05502   3.219 0.001541 **
## BMI30              -0.22160    0.07090  -3.126 0.002085 **
## Diab               -0.13744    0.05373  -2.558 0.011401 *
## Hyp_Ten            -0.19290    0.03723  -5.181 6.16e-07 ***
## Insuf_Phy           0.10295    0.02410   4.271 3.22e-05 ***
## N_HDL_Chol          2.99500    0.89305   3.354 0.000982 ***
## Pol3                0.06453    0.02283   2.827 0.005263 **
## recip_PCGMP       134.01810  622.27596   0.215 0.829737
## Tax_Ciga            4.76702    1.22382   3.895 0.000141 ***
## sqrt_Cal_Daily      0.40515    0.08955   4.524 1.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.062 on 171 degrees of freedom
## Multiple R-squared:  0.8341, Adjusted R-squared:  0.8234
## F-statistic: 78.15 on 11 and 171 DF,  p-value: < 2.2e-16
```

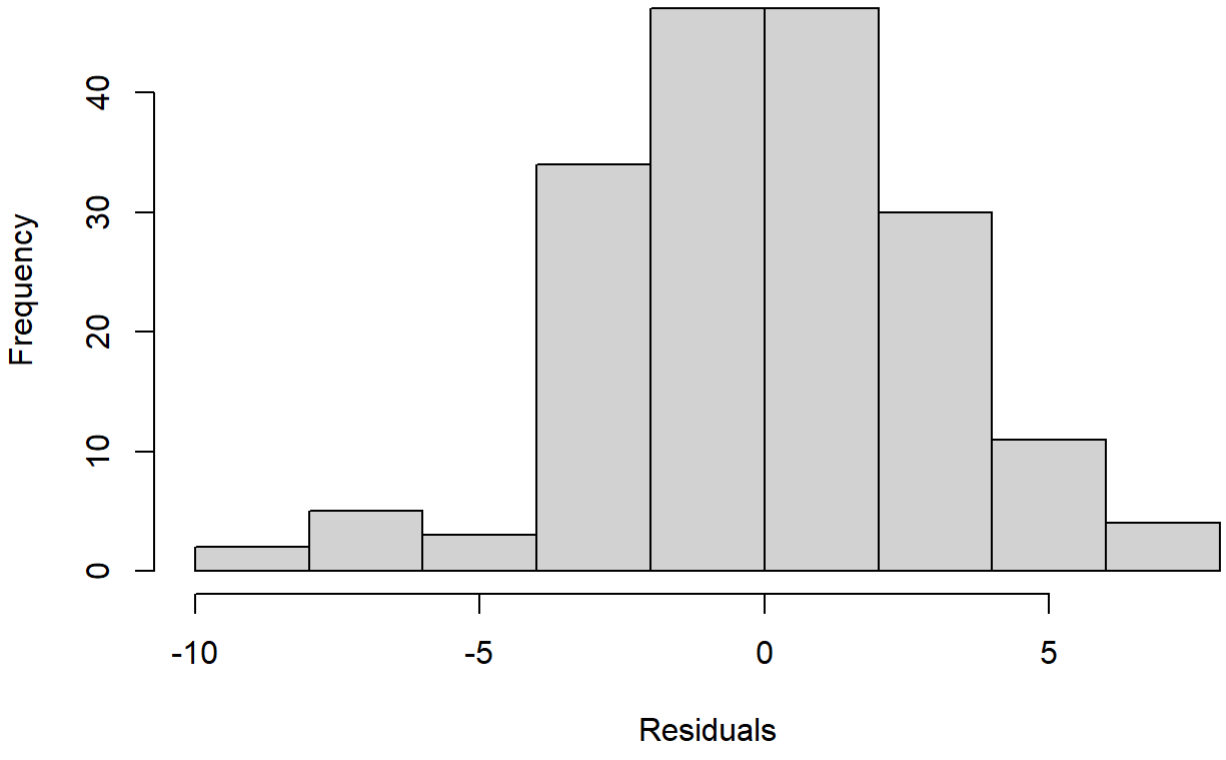
```
cat("AIC test: ", AIC(test_mod, k = 2))
```

```
## AIC test:  942.452
```

```
qqnorm(residuals(test_mod), ylab = "Residuals", main="")
qqline(residuals(test_mod))
```



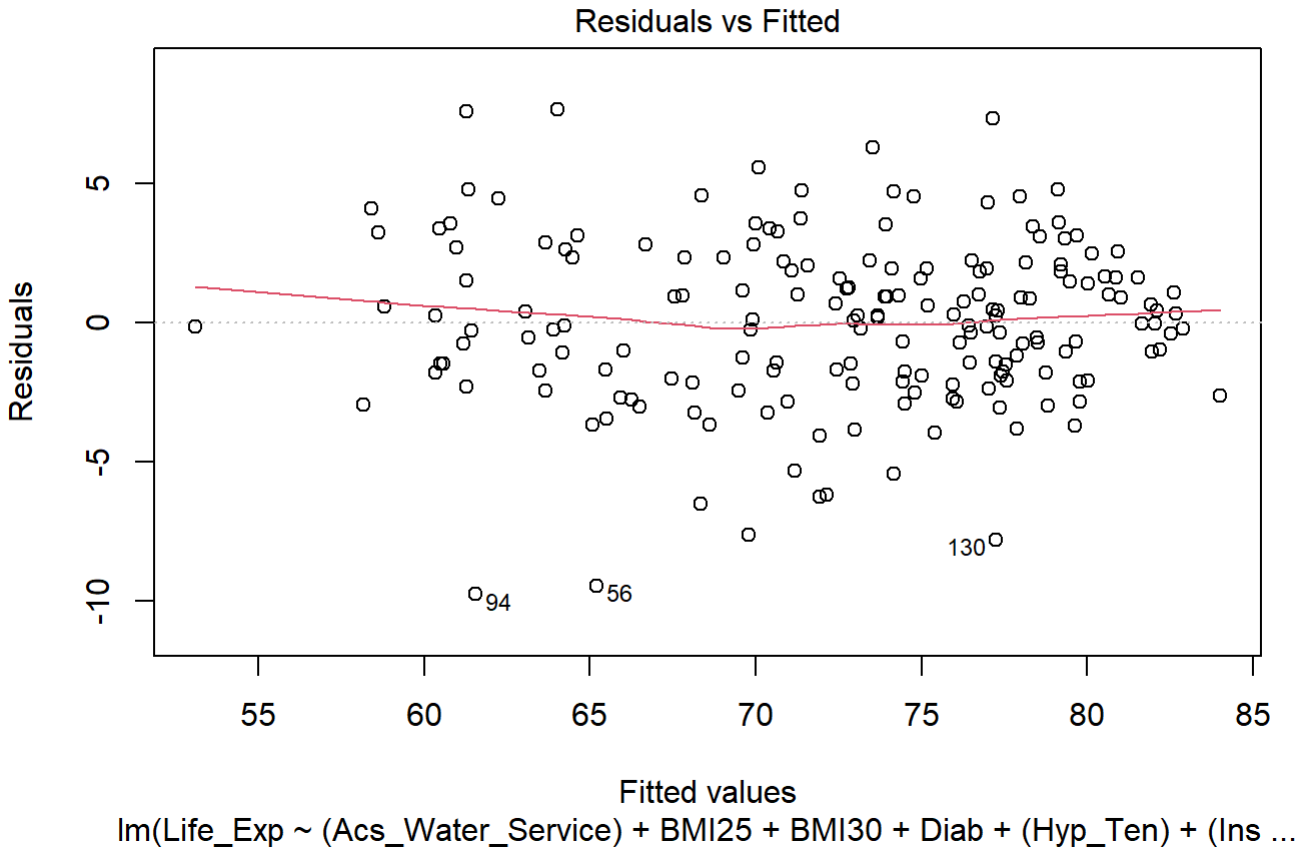

```
hist(residuals(test_mod), xlab="Residuals", main="")
```



```
shapiro.test(residuals(test_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.98554, p-value = 0.05662
```

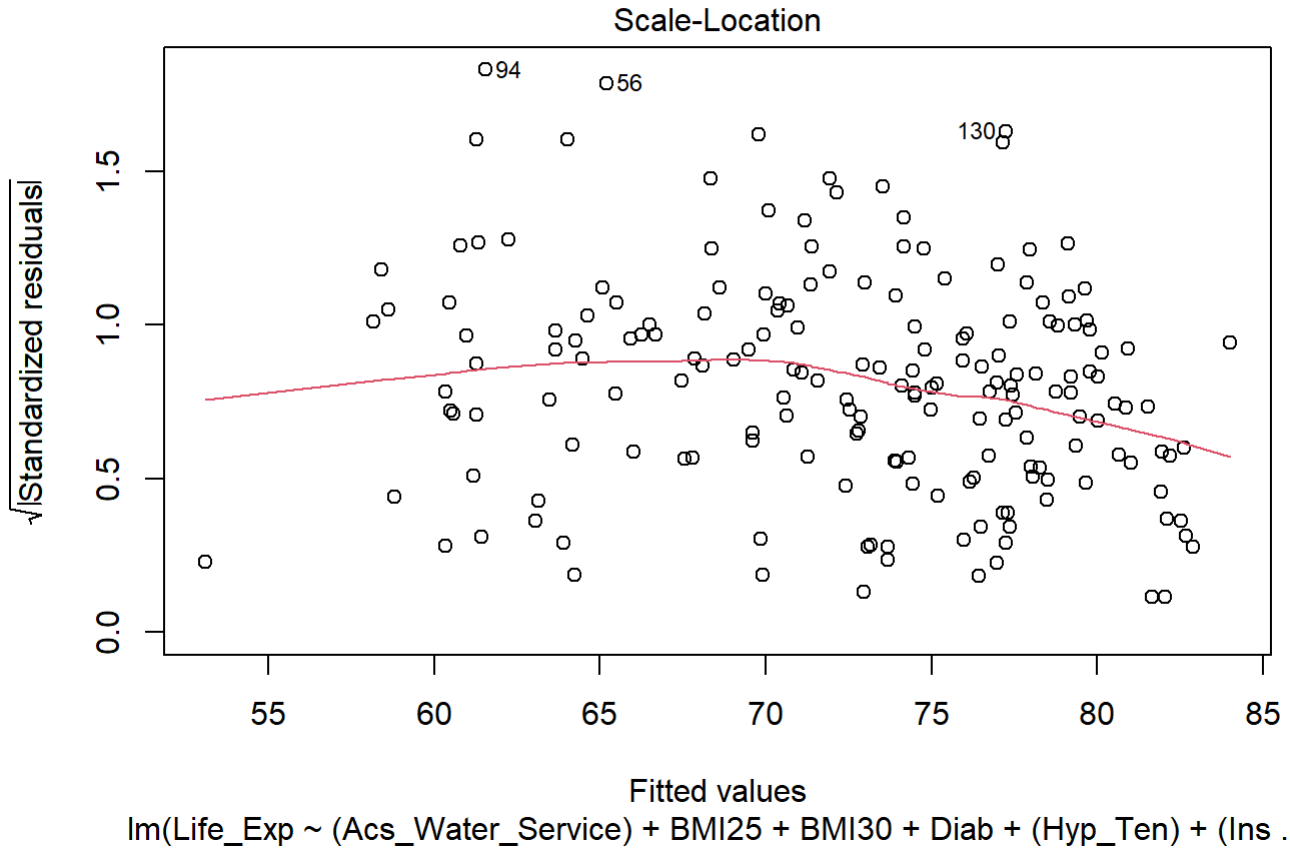
```
plot(test_mod,1)
```




```
dwtest(test_mod)
```

```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 2.0614, p-value = 0.6638
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(test_mod, 3)
```



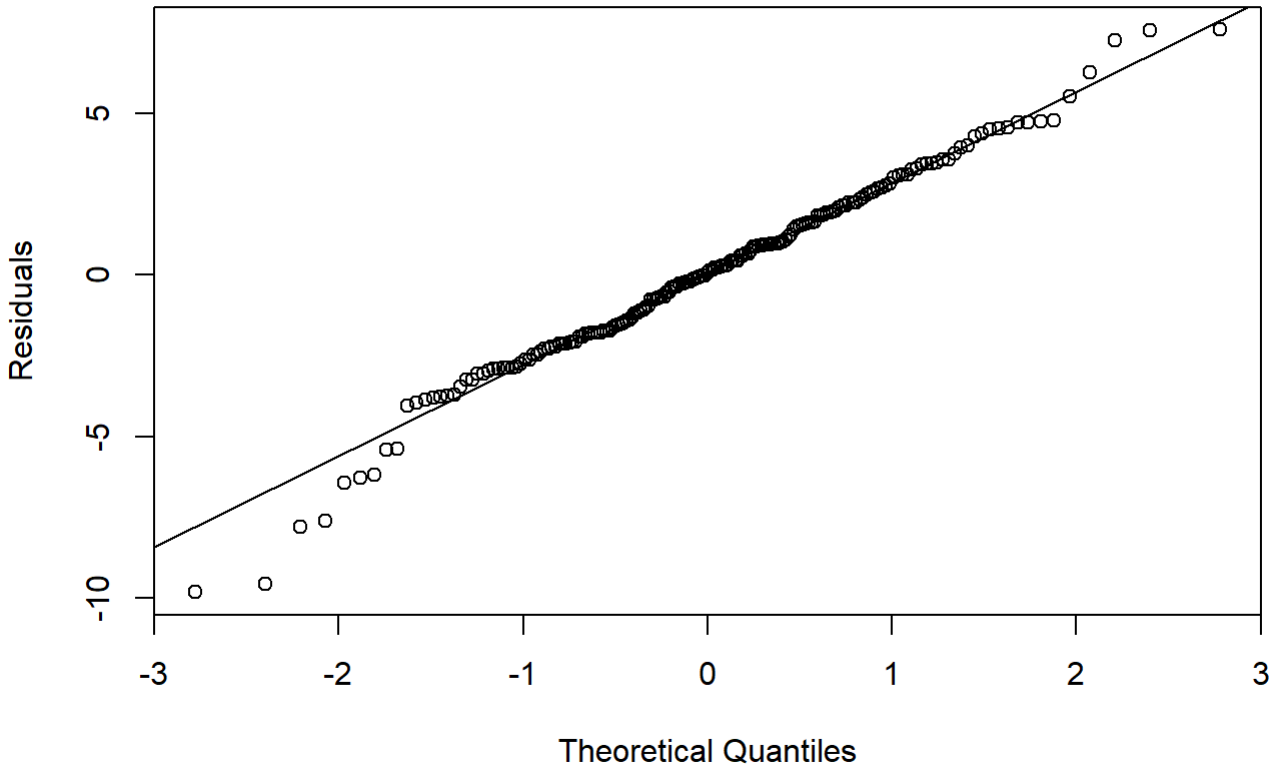

```
#sqrt(Life_Exp)

tran_BE_Imp_lmod <- lm(Life_Exp ~ (Acs_Water_Service) + BMI25 + BMI30 + Diab +
    (Hyp_Ten) + (Insuf_Phy) + (N_HDL_Chol) + (Pol3) + 1/PCGMP + (Tax_Ciga) + sqrt(Cal_Daily),
data = imp_df_184)

test_mod <- tran_BE_Imp_lmod

summary(test_mod)
```

```
##
## Call:
## lm(formula = Life_Exp ~ (Acs_Water_Service) + BMI25 + BMI30 +
##     Diab + (Hyp_Ten) + (Insuf_Phy) + (N_HDL_Chol) + (Pol3) +
##     1/PCGMP + (Tax_Ciga) + sqrt(Cal_Daily), data = imp_df_184)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8133 -1.8648  0.0584  1.9374  7.5835
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       24.27611    4.38280   5.539 1.12e-07 ***
## Acs_Water_Service  0.12341    0.02815   4.385 2.02e-05 ***
## BMI25              0.17555    0.05440   3.227 0.001497 **
## BMI30             -0.22046    0.07050  -3.127 0.002074 **
## Diab              -0.13767    0.05357  -2.570 0.011026 *
## Hyp_Ten           -0.19217    0.03698  -5.197 5.68e-07 ***
## Insuf_Phy          0.10301    0.02404   4.286 3.02e-05 ***
## N_HDL_Chol         2.96882    0.88228   3.365 0.000944 ***
## Pol3               0.06490    0.02270   2.859 0.004772 **
## Tax_Ciga           4.77932    1.21910   3.920 0.000127 ***
## sqrt(Cal_Daily)    0.39903    0.08469   4.712 5.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.053 on 172 degrees of freedom
## Multiple R-squared:  0.834,  Adjusted R-squared:  0.8244
## F-statistic: 86.44 on 10 and 172 DF,  p-value: < 2.2e-16
```

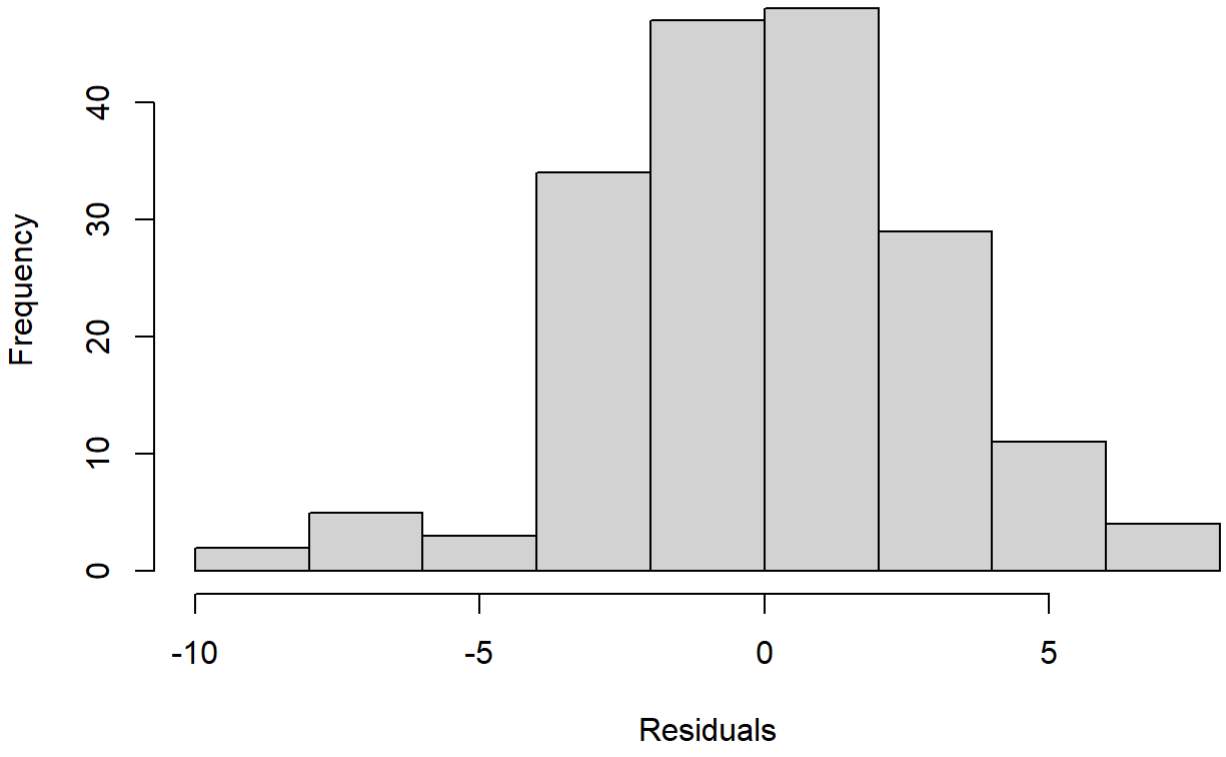
```
cat("AIC test: ", AIC(test_mod, k = 2))
```

```
## AIC test:  940.5016
```

```
qqnorm(residuals(test_mod), ylab = "Residuals", main="")
qqline(residuals(test_mod))
```



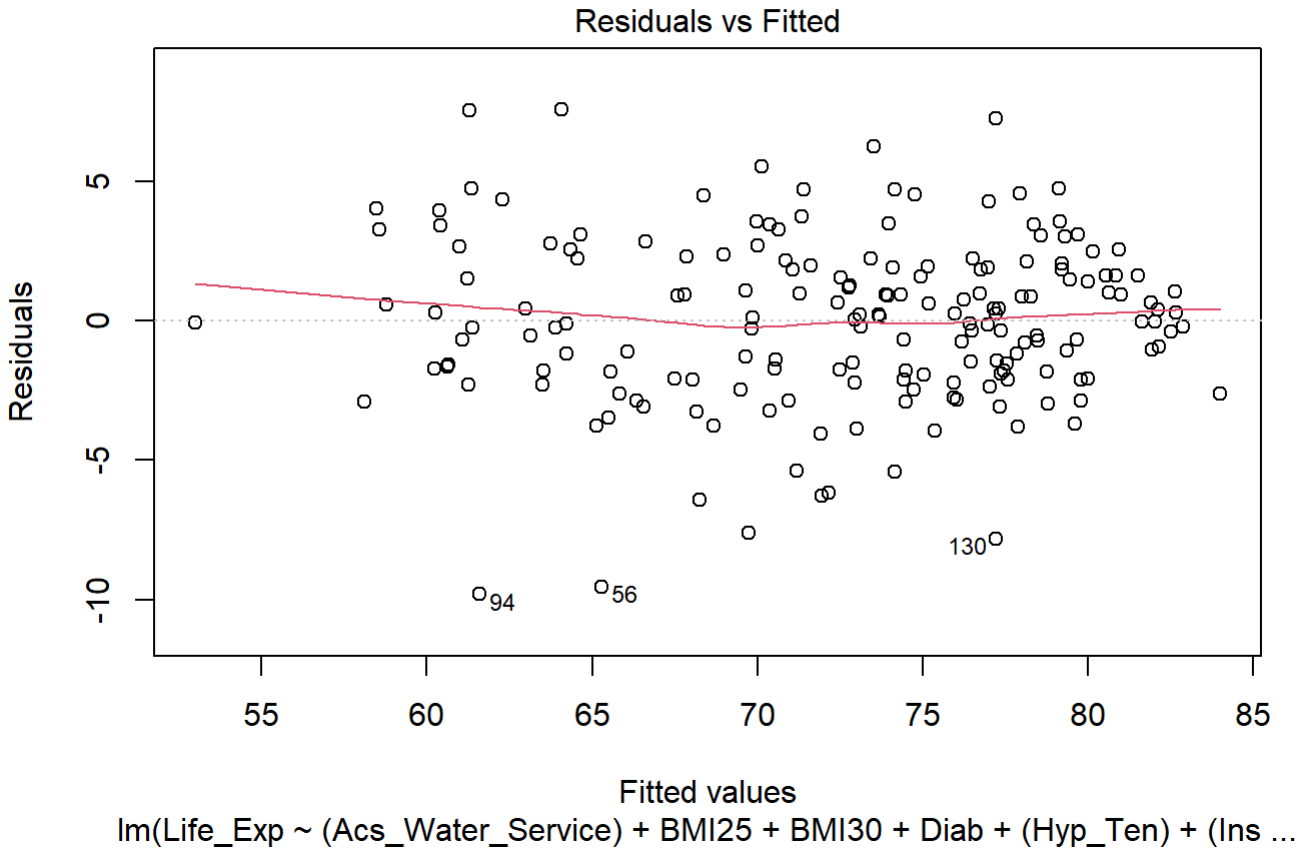

```
hist(residuals(test_mod), xlab="Residuals", main="")
```



```
shapiro.test(residuals(test_mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(test_mod)
## W = 0.98522, p-value = 0.05107
```

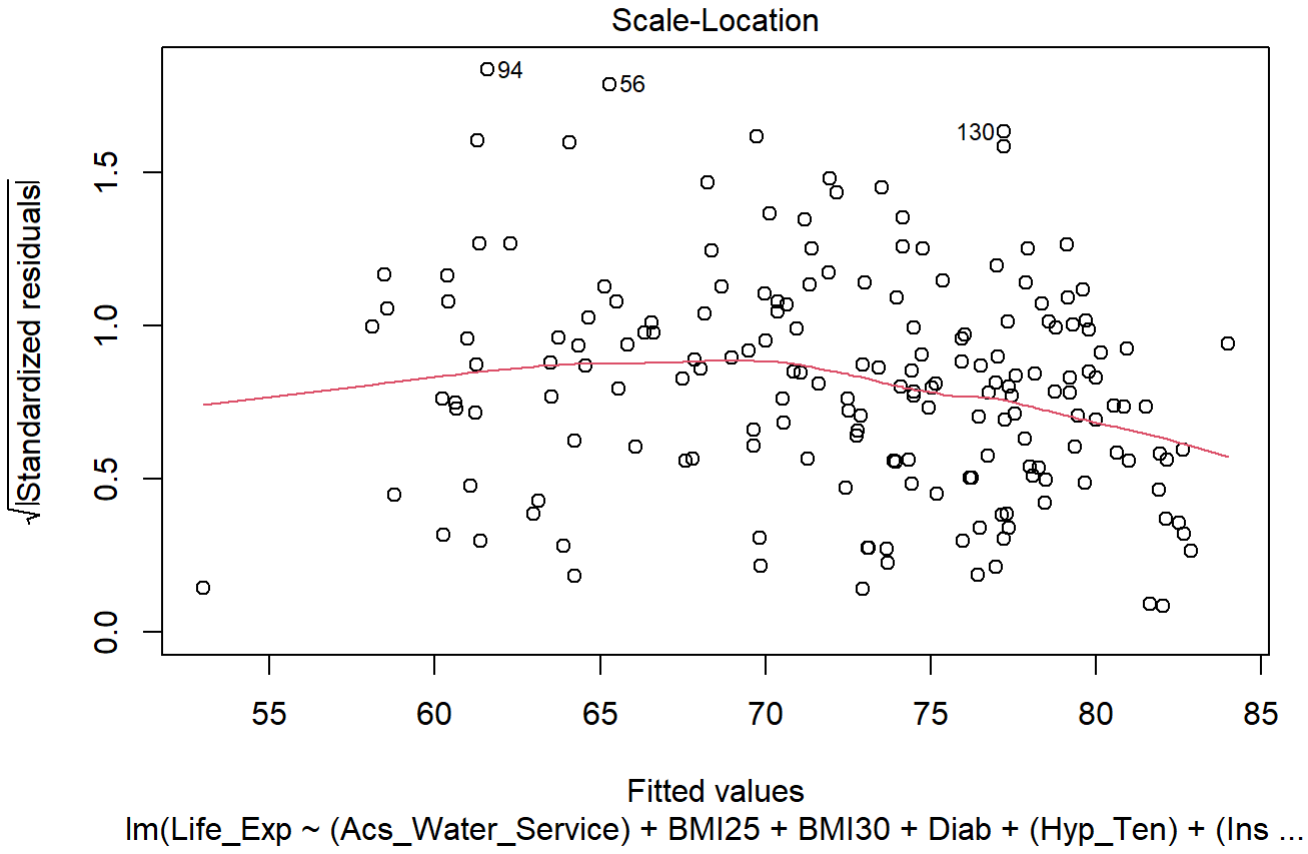
```
plot(test_mod,1)
```




```
dwtest(test_mod)
```

```
##
##  Durbin-Watson test
##
## data:  test_mod
## DW = 2.0614, p-value = 0.6659
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(test_mod, 3)
```


Scale-Location
94
56
130
√|Standardized residuals|
0.0
0.5
1.0
1.5
55
60
65
70
75
80
85
Fitted values
lm(Life_Exp ~ (Acs_Water_Service) + BMI25 + BMI30 + Diab + (Hyp_Ten) + (Ins ...