Final project by Ian Musumba

2022-10-29

#Quick view of the datset

```
glimpse(life)
```

```
## Rows: 2,938
## Columns: 22
## $ Country                          <chr> "Afghanistan", "Afghanistan", "Afghani~
## $ Year                             <int> 2015, 2014, 2013, 2012, 2011, 2010, 20~
## $ Status                           <chr> "Developing", "Developing", "Developin~
## $ Life.expectancy                  <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, 58~
## $ Adult.Mortality                  <int> 263, 271, 268, 272, 275, 279, 281, 287~
## $ infant.deaths                    <int> 62, 64, 66, 69, 71, 74, 77, 80, 82, 84~
## $ Alcohol                          <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.~
## $ percentage.expenditure           <dbl> 71.279624, 73.523582, 73.219243, 78.18~
## $ Hepatitis.B                      <int> 65, 62, 64, 67, 68, 66, 63, 64, 63, 64~
## $ Measles                          <int> 1154, 492, 430, 2787, 3013, 1989, 2861~
## $ BMI                              <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, 16~
## $ under.five.deaths                <int> 83, 86, 89, 93, 97, 102, 106, 110, 113~
## $ Polio                            <int> 6, 58, 62, 67, 68, 66, 63, 64, 63, 58,~
## $ Total.expenditure                <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, 9.~
## $ Diphtheria                       <int> 65, 62, 64, 67, 68, 66, 63, 64, 63, 58~
## $ HIV.AIDS                         <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1~
## $ GDP                              <dbl> 584.25921, 612.69651, 631.74498, 669.9~
## $ Population                       <dbl> 33736494, 327582, 31731688, 3696958, 2~
## $ thinness..1.19.years             <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, 18~
## $ thinness.5.9.years               <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, 18~
## $ Income.composition.of.resources <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0.4~
## $ Schooling                        <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9, 8~
```

Checking for null values

```
table(is.na(life))
```

```
##
## FALSE  TRUE
## 62073  2563
```

Handling missing values by replacing them with median value

```
life$Life.expectancy[is.na(life$Life.expectancy)] <- median(life$Life.expectancy, na.rm = T)
life$Schooling[is.na(life$Schooling)] <- median(life$Schooling, na.rm = T)
life$infant.deaths[is.na(life$infant.deaths)] <- median(life$infant.deaths, na.rm = T)
life$Hepatitis.B[is.na(life$Hepatitis.B)] <- median(life$Hepatitis.B, na.rm = T)
life$BMI[is.na(life$BMI)] <- median(life$BMI, na.rm = T)
life$GDP[is.na(life$GDP)] <- median(life$GDP, na.rm = T)
life$Population[is.na(life$Population)] <- median(life$Population, na.rm = T)
life$Income.composition.of.resources[is.na(life$Income.composition.of.resources)] <-   median(life$Incor
```

```
cor(life[,c(4,  6, 9, 11, 16:18, 21:22)], use = "complete.obs")
```

```
##                                 Life.expectancy infant.deaths Hepatitis.B
## Life.expectancy                      1.00000000   -0.19676906  0.17021864
## infant.deaths                       -0.19676906    1.00000000 -0.16742088
## Hepatitis.B                          0.17021864   -0.16742088  1.00000000
## BMI                                  0.55690117   -0.22679646  0.11244122
## HIV.AIDS                            -0.55670342    0.02523132 -0.08549672
## GDP                                  0.43046130   -0.10282895  0.07665968
## Population                          -0.02901388    0.55166746 -0.12500551
## Income.composition.of.resources      0.68866162   -0.14157131  0.11765158
## Schooling                            0.71305353   -0.19095097  0.14127478
##                                          BMI    HIV.AIDS         GDP  Population
## Life.expectancy                   0.55690117 -0.55670342  0.43046130 -0.02901388
## infant.deaths                    -0.22679646  0.02523132 -0.10282895  0.55166746
## Hepatitis.B                       0.11244122 -0.08549672  0.07665968 -0.12500551
## BMI                               1.00000000 -0.24338267  0.27393222 -0.06966749
## HIV.AIDS                         -0.24338267  1.00000000 -0.12258994 -0.01709429
## GDP                               0.27393222 -0.12258994  1.00000000 -0.02526882
## Population                       -0.06966749 -0.01709429 -0.02526882  1.00000000
## Income.composition.of.resources   0.47194664 -0.24782302  0.43595983 -0.01723712
## Schooling                         0.49980620 -0.21882240  0.43222866 -0.03681369
##                                 Income.composition.of.resources   Schooling
## Life.expectancy                                      0.68866162  0.71305353
## infant.deaths                                       -0.14157131 -0.19095097
## Hepatitis.B                                          0.11765158  0.14127478
## BMI                                                  0.47194664  0.49980620
## HIV.AIDS                                            -0.24782302 -0.21882240
## GDP                                                  0.43595983  0.43222866
## Population                                          -0.01723712 -0.03681369
## Income.composition.of.resources                      1.00000000  0.79538328
## Schooling                                            0.79538328  1.00000000
```

#EDA

#Summary statistics

```
favstats(~Life.expectancy, data = life)
```

```
##   min   Q1 median   Q3 max     mean       sd    n missing
##  36.3 63.2   72.1 75.6  89 69.23472 9.509115 2938       0
```

```
favstats(~BMI, data = life)
```

```
##  min   Q1 median   Q3  max     mean       sd    n missing
##    1 19.4   43.5 56.1 87.3 38.38118 19.93537 2938       0
```

```
favstats(~GDP, data = life)
```

```
##       min       Q1   median       Q3      max     mean       sd    n missing
##  1.68135 580.487 1766.948 4779.405 119172.7 6611.524 13296.6 2938       0
```

#Histogram of Life Expectancy

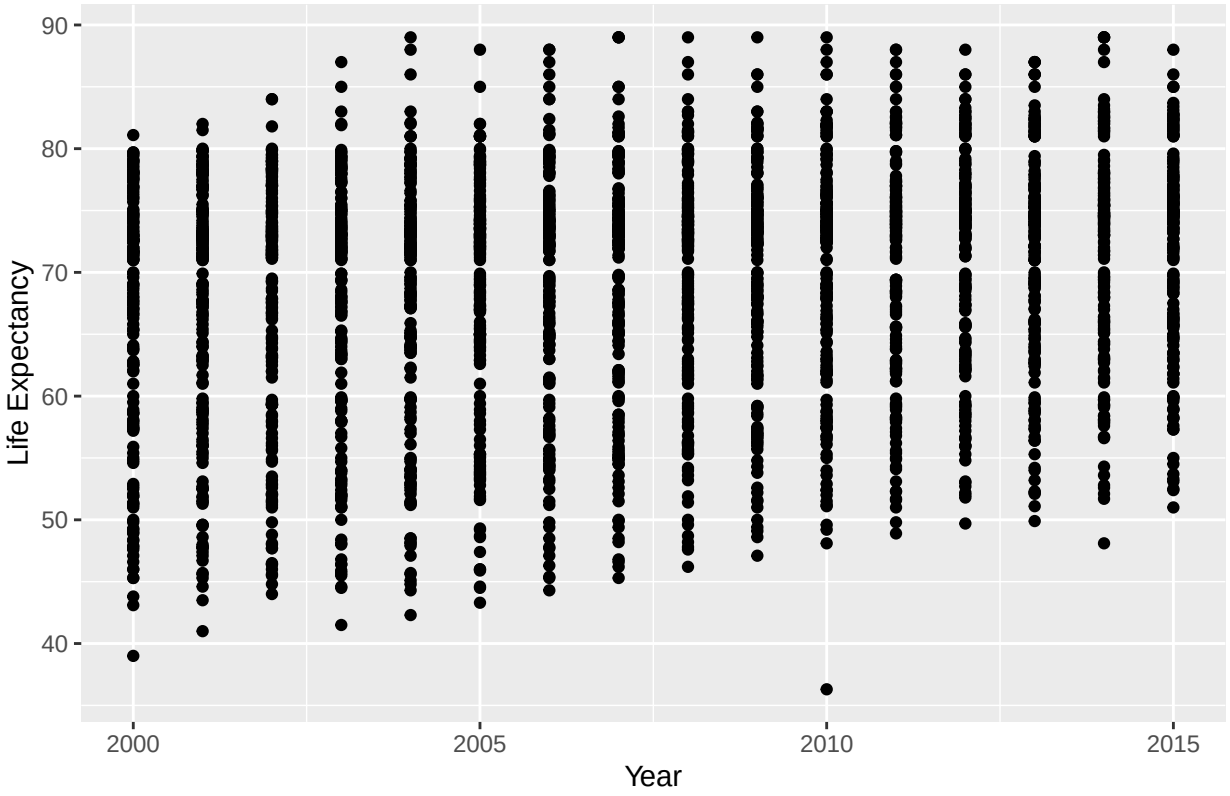
```
ggplot(life, aes(x=Life.expectancy)) + geom_density(col="blue") +
  geom_histogram(aes(y=..density..), colour="black", fill=NA) + ggtitle("Figure 1: Distribution of Life
```

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

#Scatter plot of Years vs. Life expectancy

```
ggplot(life, aes(x=Year, y=Life.expectancy)) + geom_point() + ggtitle("Figure 2: Year vs Life Expectanc
  xlab("Year") + ylab("Life Expectancy")
```

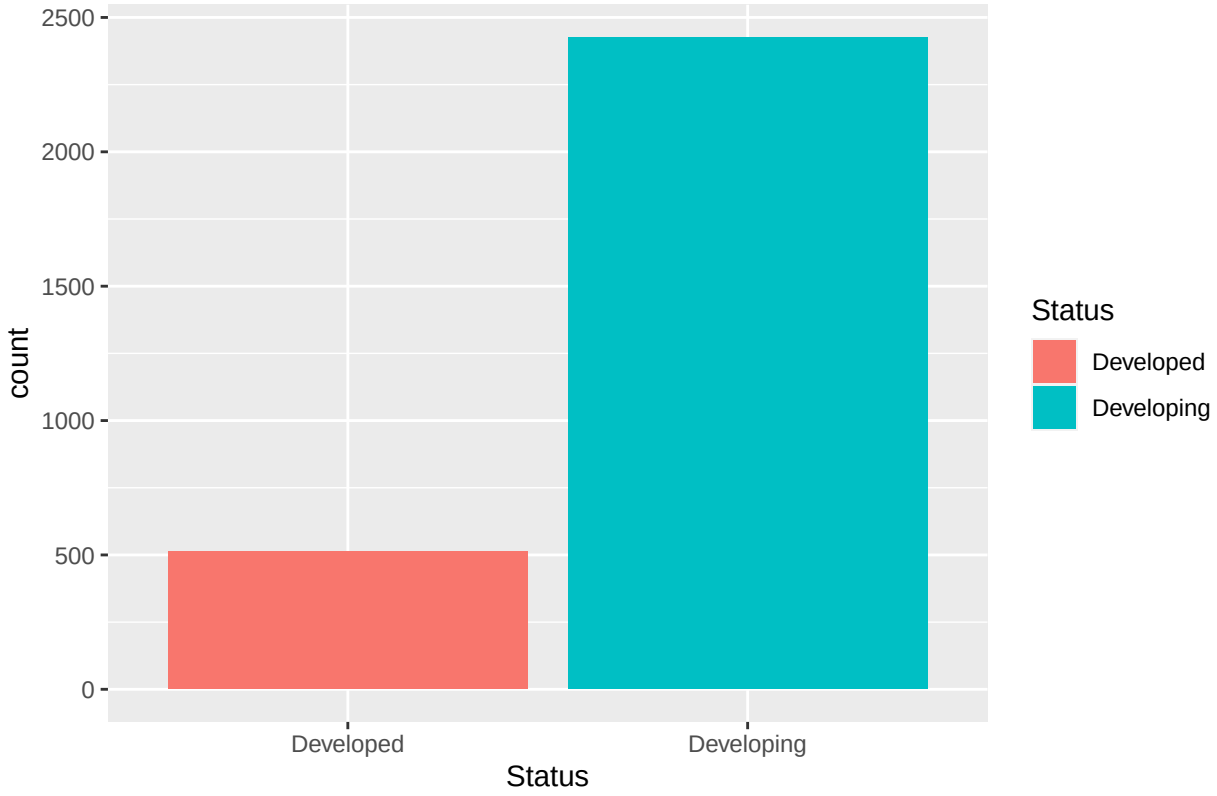
Figure 2: Year vs Life Expectancy



#One categorical variable barchart

```
ggplot(life, aes(x=Status, fill=Status)) + geom_bar(position="dodge") + ggtitle("Figure 3: Bar chart of
```

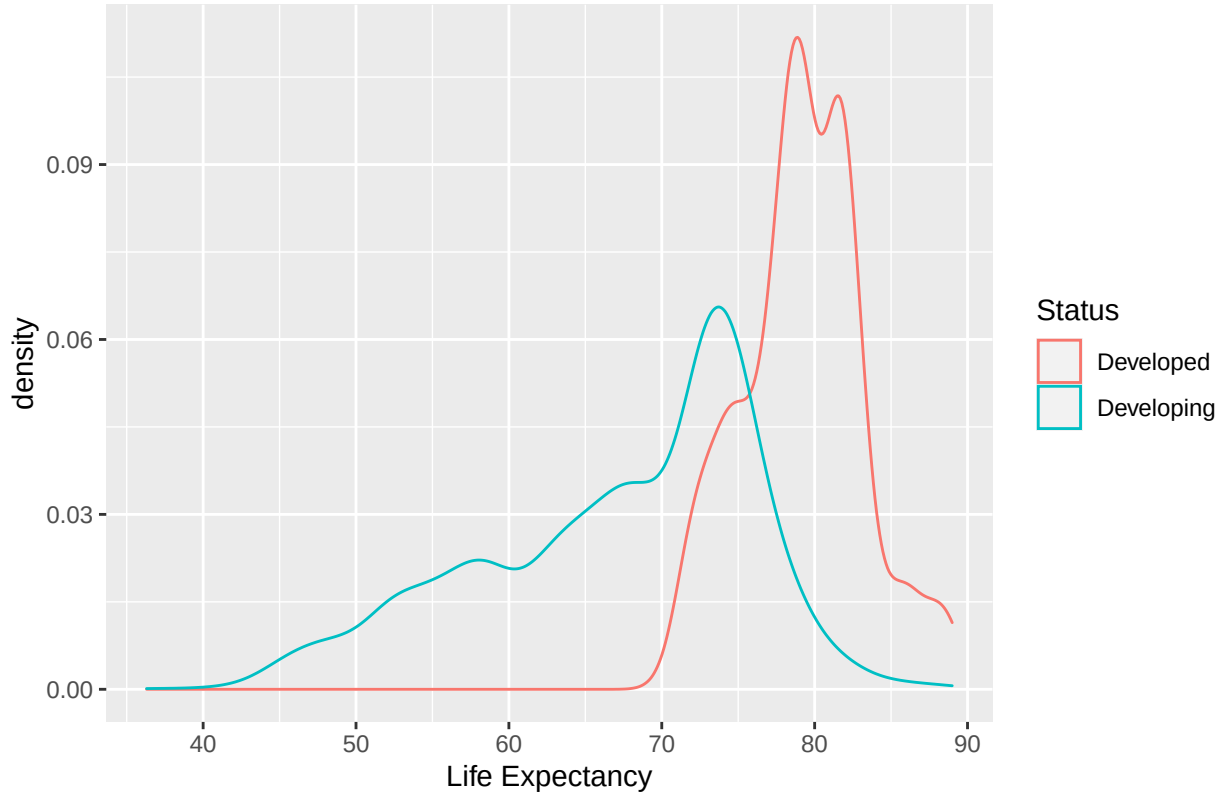
Figure 3: Bar chart of Status



#Checking whether Status has an effect on Life expactancy

```
ggplot(life, aes(x=Life.expectancy, col=Status)) + geom_density() + ggtitle("Figure 4: Overlaid Density
```

Figure 4: Overlaid Density Plots of Life Expectancy by Status



#fitting models

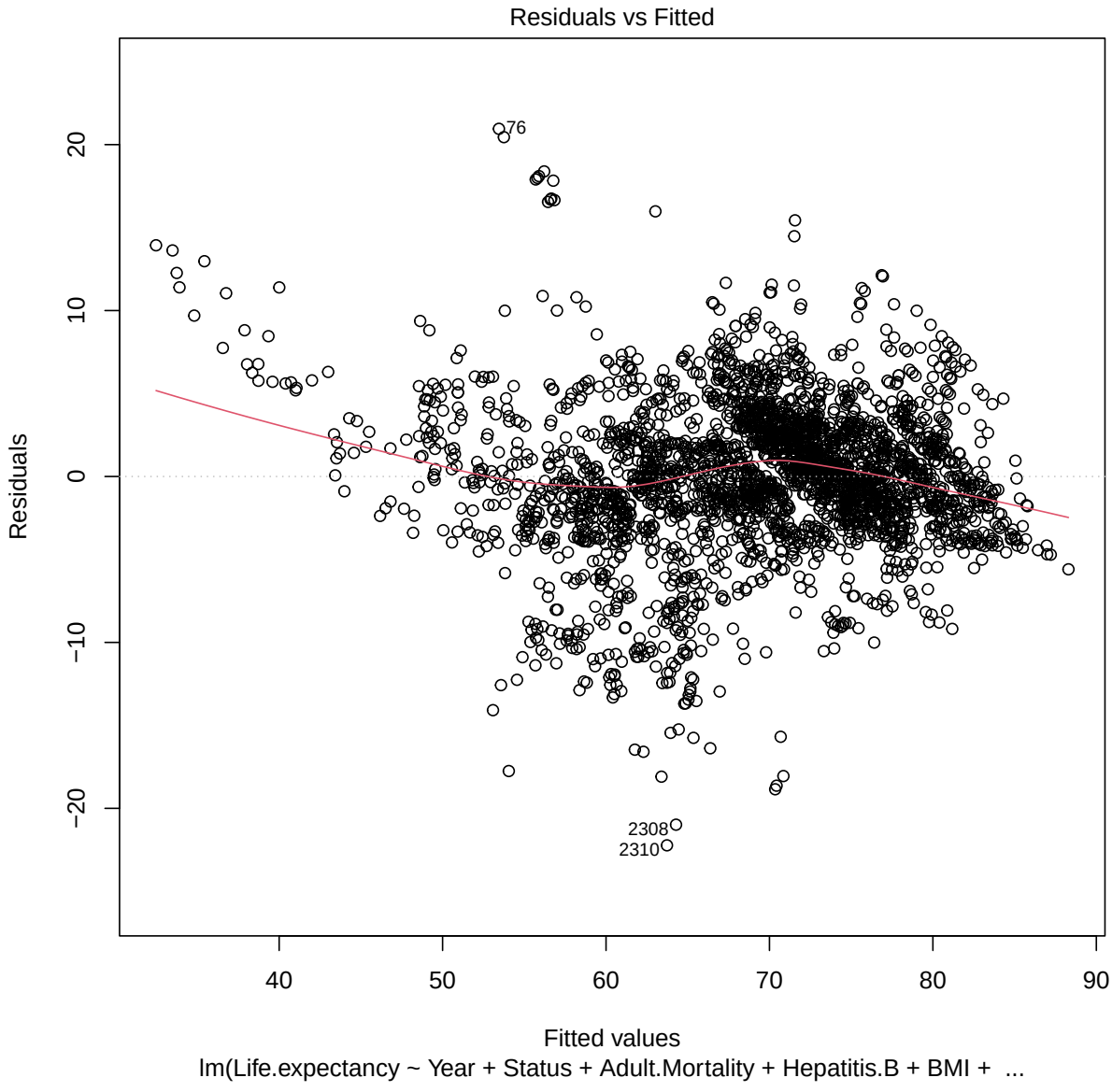
Model 1:

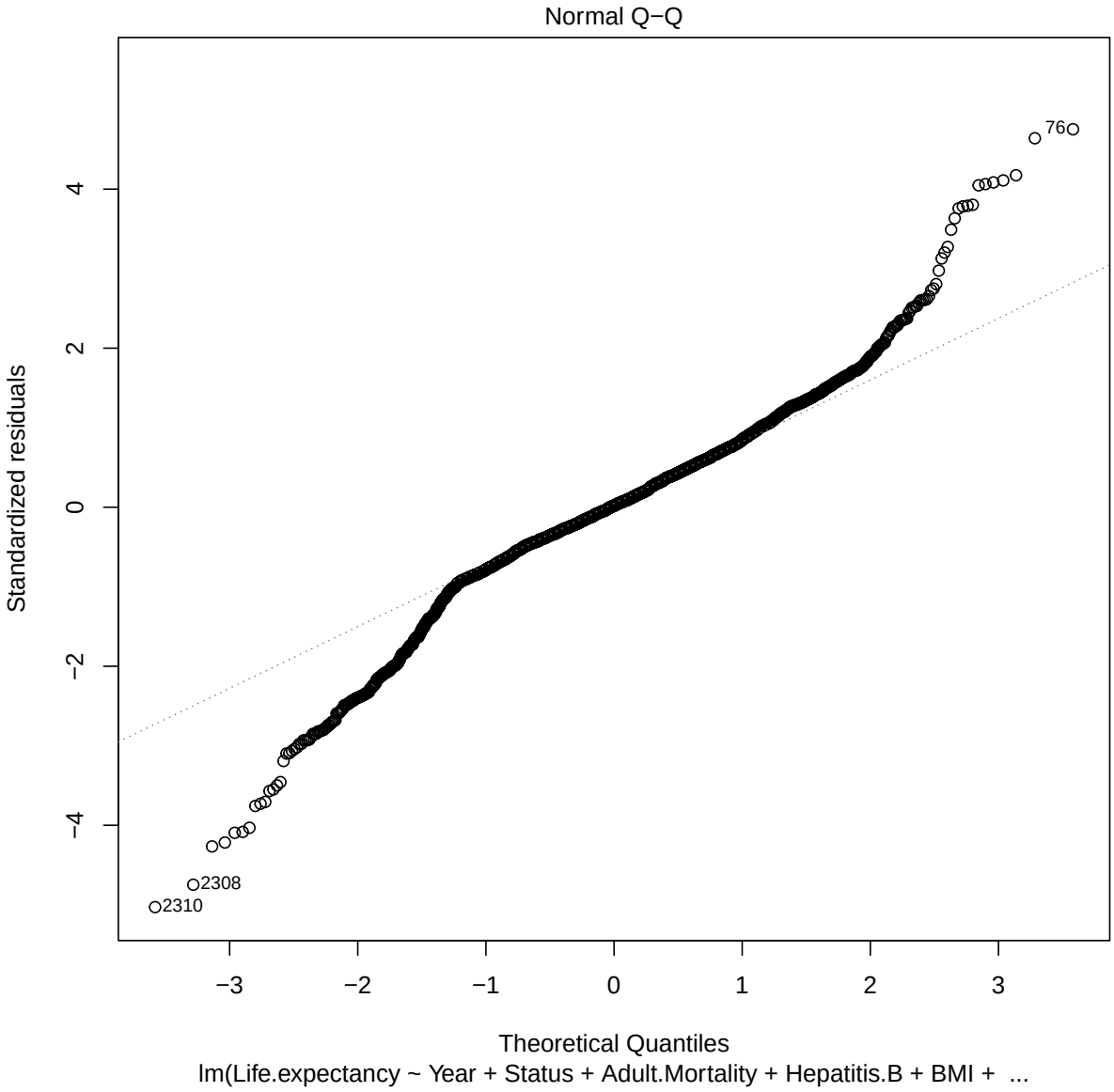
```
MODEL1 <- lm(Life.expectancy ~ Year+ Status+ Adult.Mortality+ Hepatitis.B+ BMI+ GDP+ Population+ Income
summary(MODEL1)
```

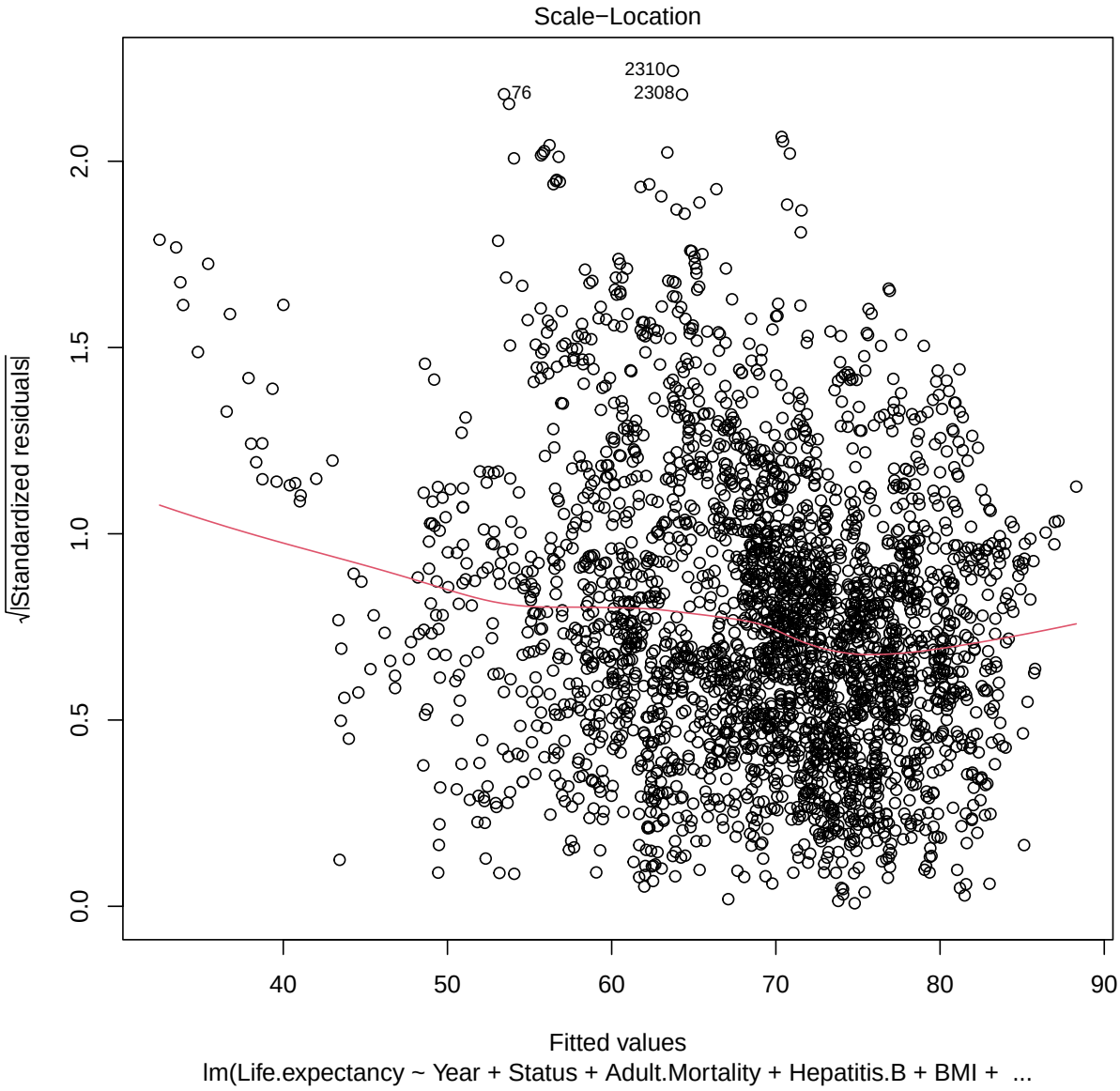
```
##
## Call:
## lm(formula = Life.expectancy ~ Year + Status + Adult.Mortality +
##     Hepatitis.B + BMI + GDP + Population + Income.composition.of.resources +
##     HIV.AIDS + Schooling, data = life, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.2291  -2.0947   0.0782   2.5287  20.9528
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.631e+01  3.716e+01   1.784  0.07449 .
## Year                           -4.643e-03  1.857e-02  -0.250  0.80263
## StatusDeveloping               -1.848e+00  2.643e-01  -6.990 3.39e-12 ***
## Adult.Mortality                -2.127e-02  8.605e-04 -24.713  < 2e-16 ***
## Hepatitis.B                     1.189e-02  3.646e-03   3.262  0.00112 **
## BMI                             6.215e-02  4.947e-03  12.563  < 2e-16 ***
## GDP                             4.306e-05  7.226e-06   5.959 2.85e-09 ***
```

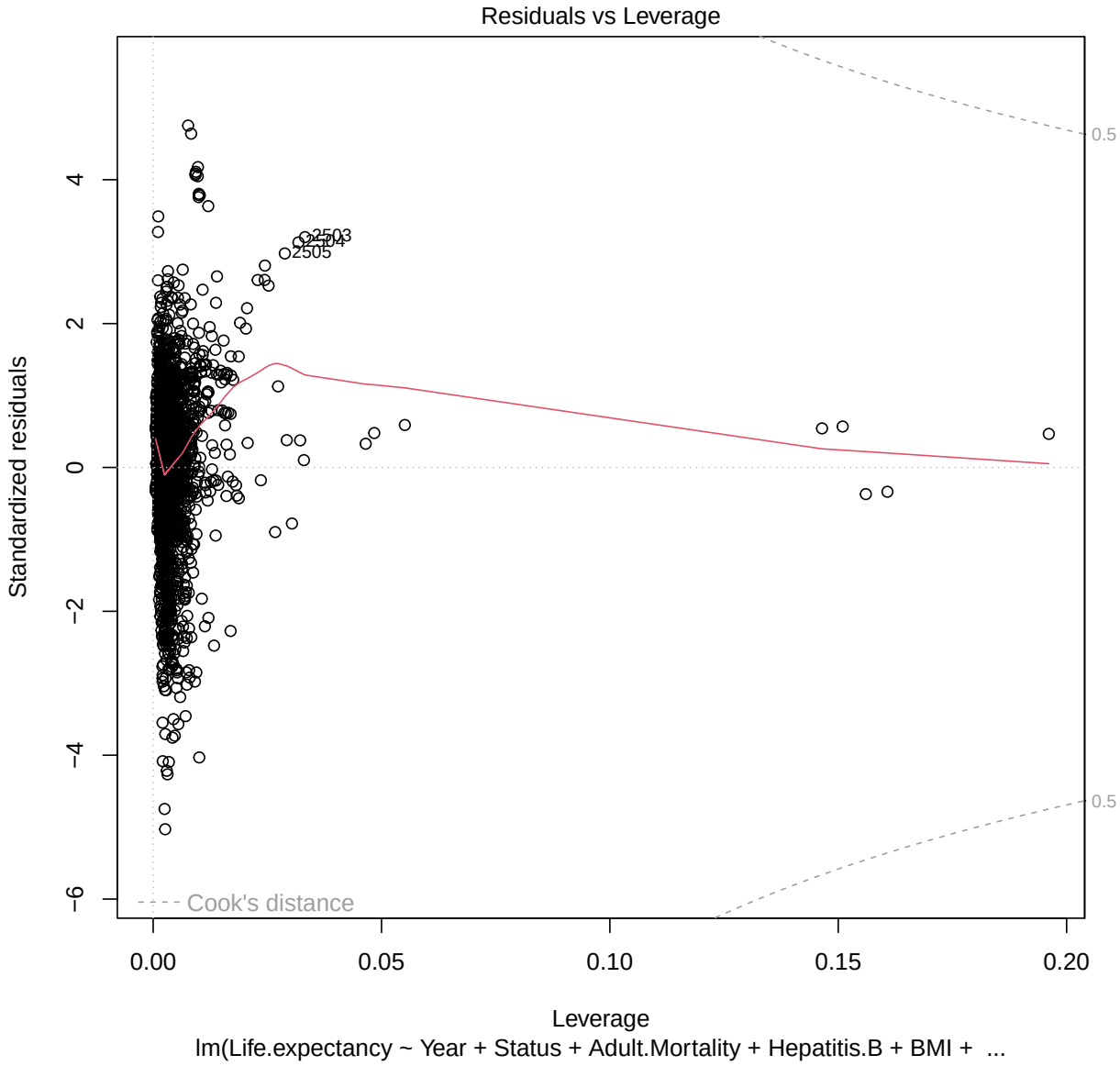
```
## Population                      -8.061e-10  1.529e-09  -0.527  0.59813
## Income.composition.of.resources  6.869e+00  6.910e-01   9.941  < 2e-16 ***
## HIV.AIDS                        -4.848e-01  1.909e-02 -25.391  < 2e-16 ***
## Schooling                        8.440e-01  4.477e-02  18.854  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.426 on 2917 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.7848, Adjusted R-squared:  0.7841
## F-statistic:  1064 on 10 and 2917 DF,  p-value: < 2.2e-16
```

```
plot(MODEL1)
```

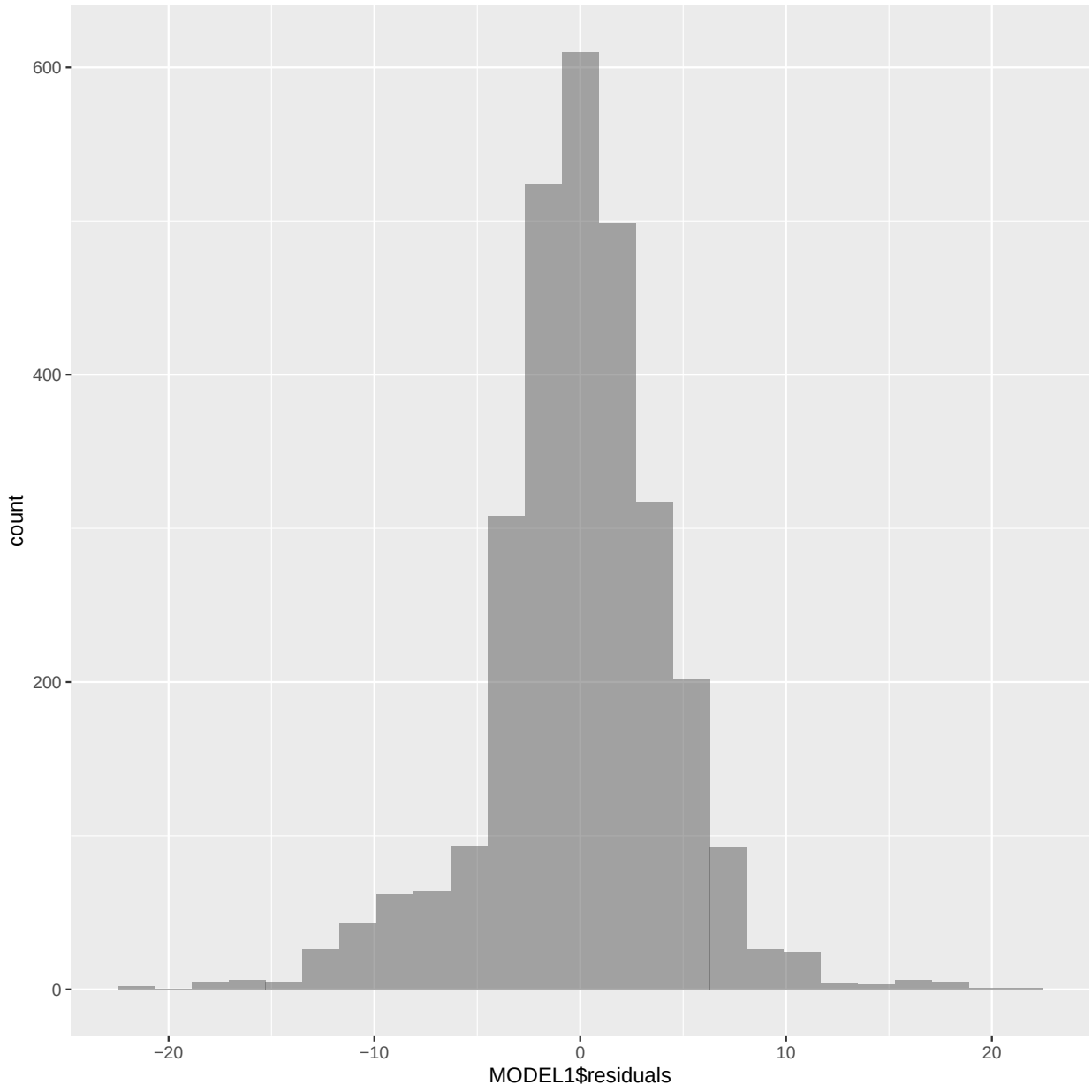



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Life.expectancy ~ Year + Status + Adult.Mortality + Hepatitis.B + BMI + ...
76
2308
2310
−4
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
2310
2308
76
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
40
50
60
70
80
90
Fitted values
lm(Life.expectancy ~ Year + Status + Adult.Mortality + Hepatitis.B + BMI + ...




```
gf_histogram(~MODEL1$residuals)
```




```
mean(MODEL1$residual^2)
```

```
## [1] 19.51321
```

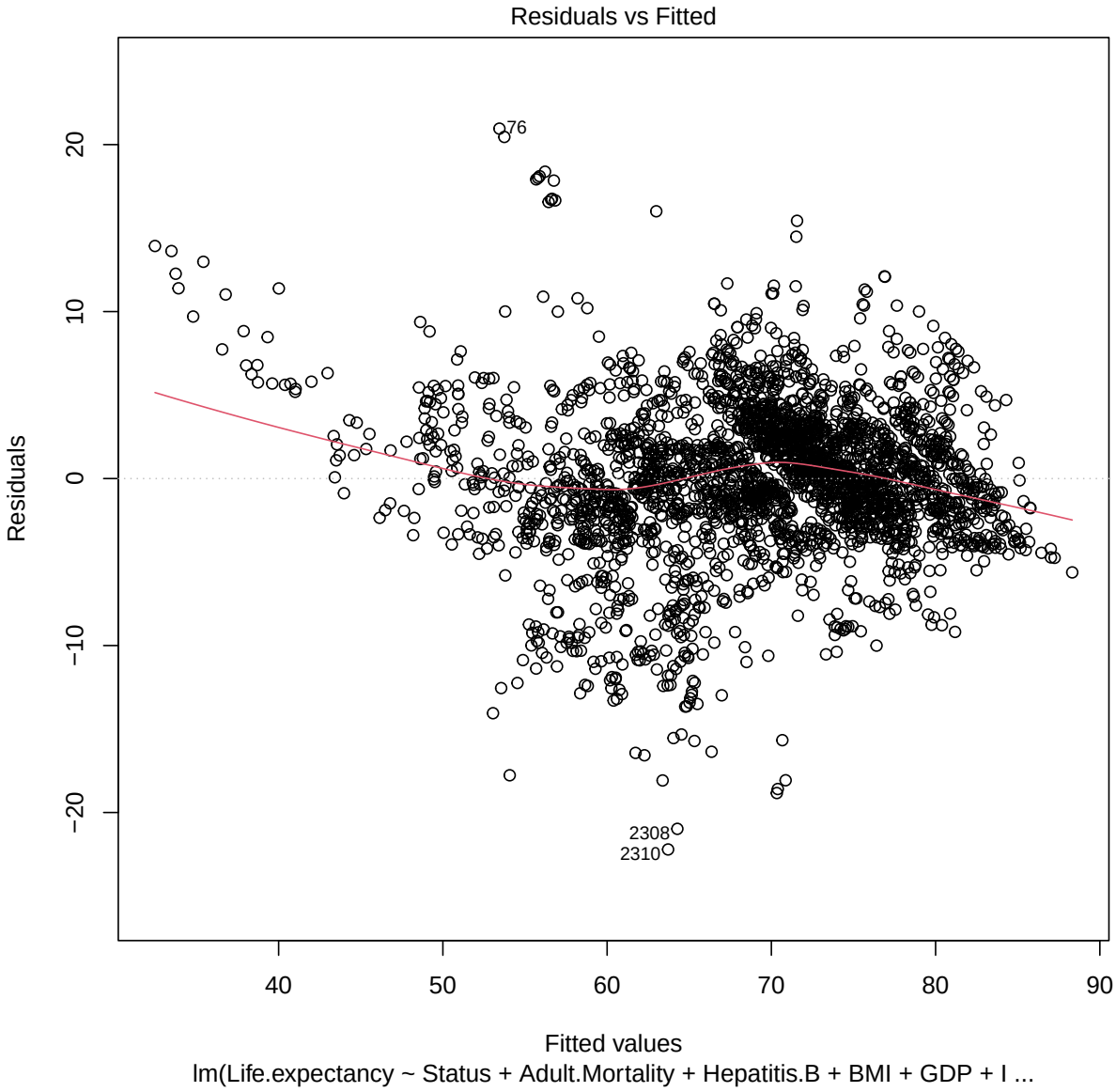
Final Model:

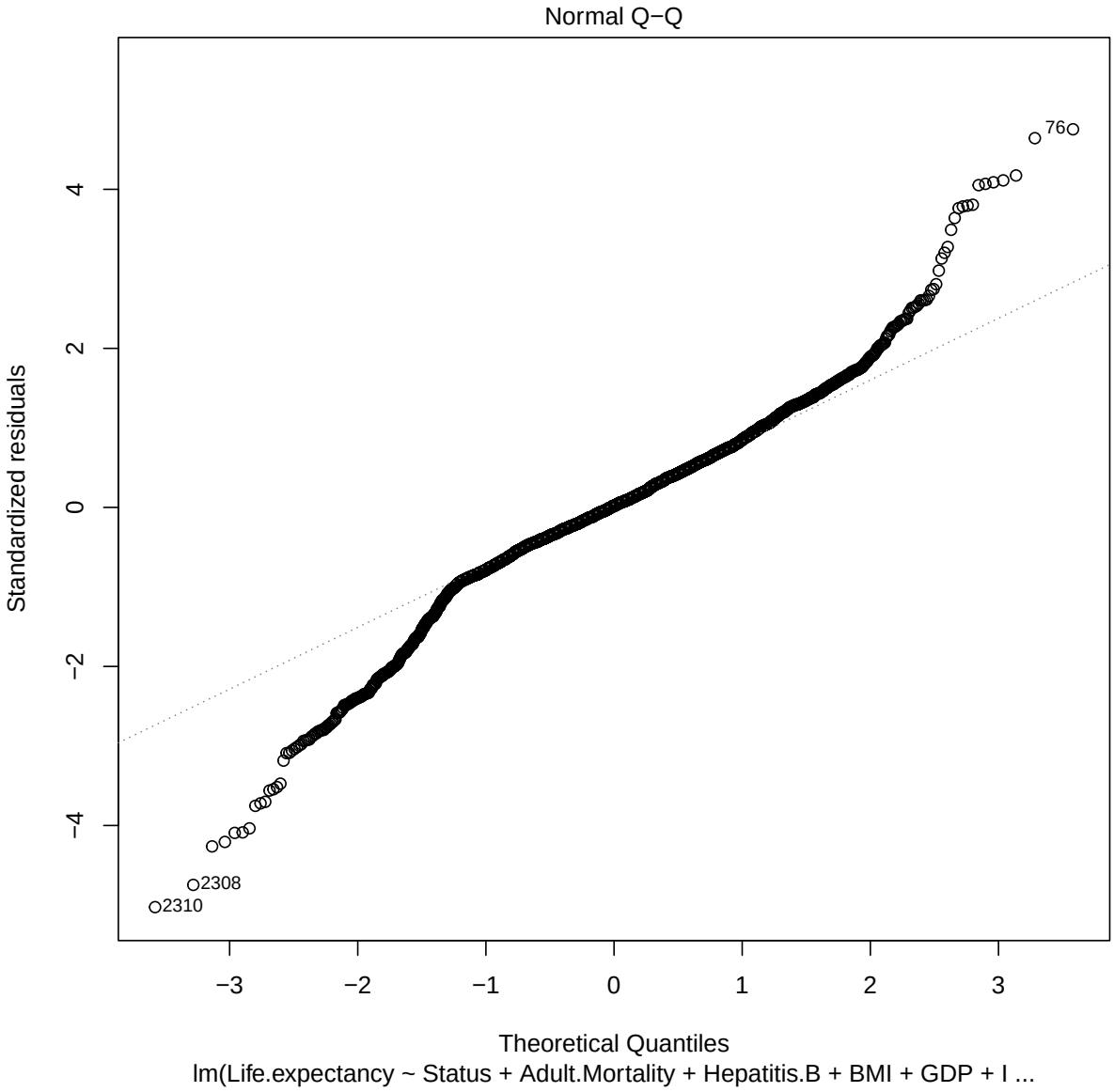
```
MODEL2 <- lm(Life.expectancy ~ Status+ Adult.Mortality+ Hepatitis.B+ BMI+ GDP+ Income.composition.of.res
summary(MODEL2)
```

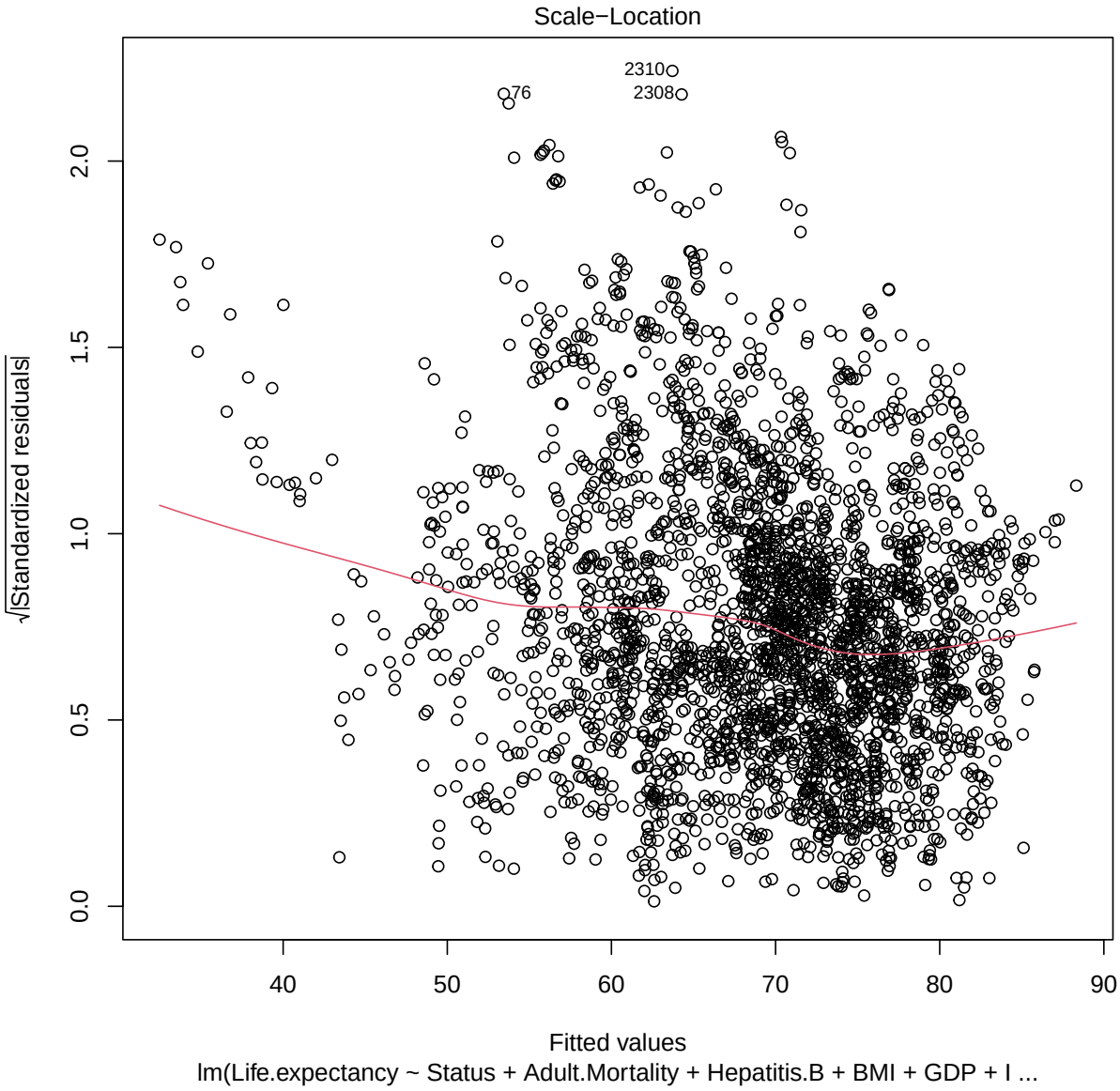
```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + Hepatitis.B +
```

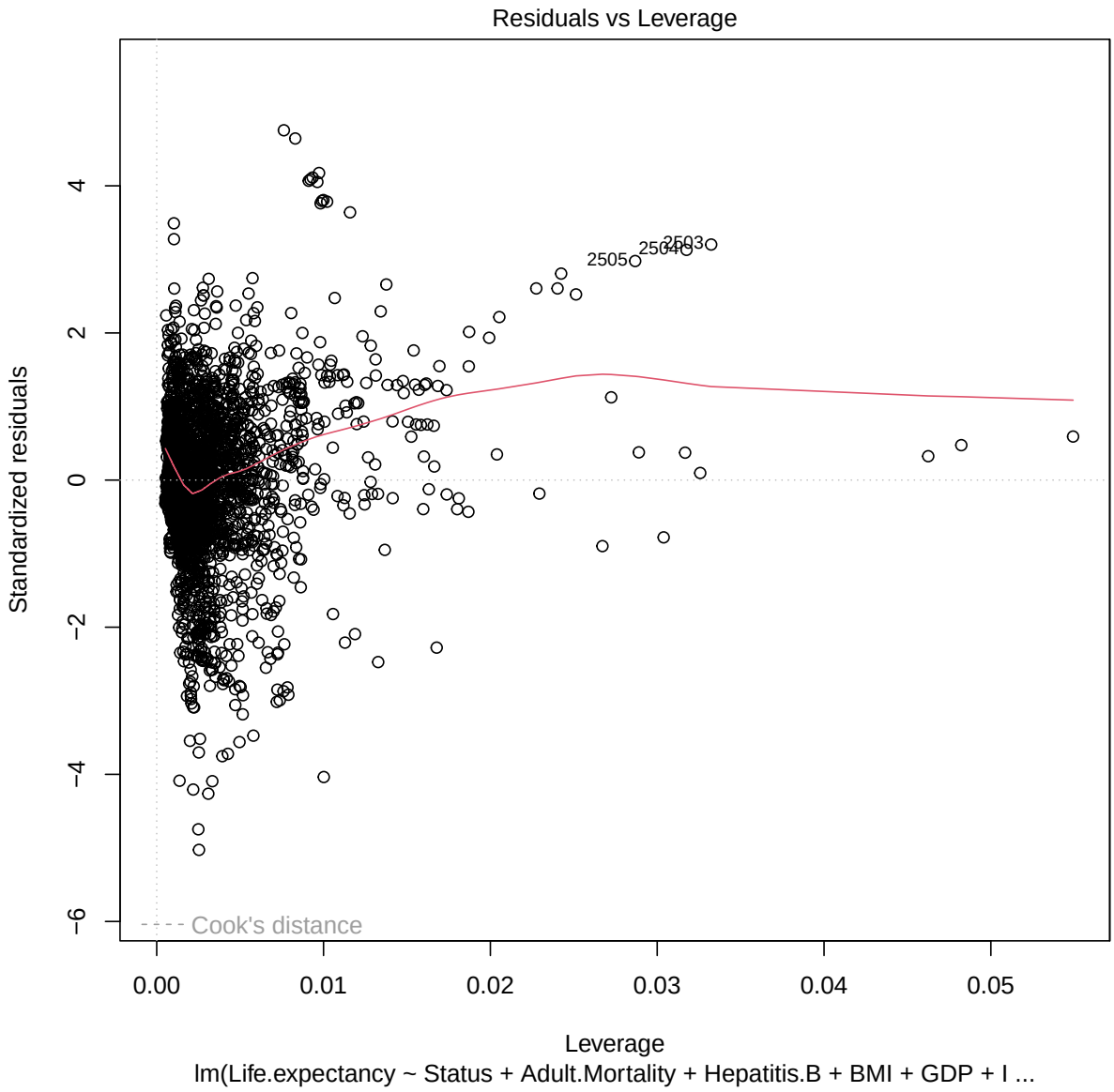
```
##     BMI + GDP + Income.composition.of.resources + HIV.AIDS +
##     Schooling, data = life, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.2128  -2.1111   0.0863   2.5244  20.9571
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.699e+01  6.258e-01  91.053  < 2e-16 ***
## StatusDeveloping               -1.859e+00  2.618e-01  -7.099 1.57e-12 ***
## Adult.Mortality                -2.127e-02  8.579e-04 -24.792  < 2e-16 ***
## Hepatitis.B                     1.213e-02  3.618e-03   3.352 0.000813 ***
## BMI                             6.233e-02  4.936e-03  12.628  < 2e-16 ***
## GDP                             4.303e-05  7.221e-06   5.960 2.83e-09 ***
## Income.composition.of.resources  6.837e+00  6.848e-01   9.985  < 2e-16 ***
## HIV.AIDS                       -4.840e-01  1.896e-02 -25.528  < 2e-16 ***
## Schooling                       8.436e-01  4.464e-02  18.898  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.424 on 2919 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.7848, Adjusted R-squared:  0.7842
## F-statistic:  1330 on 8 and 2919 DF,  p-value: < 2.2e-16
```

```
plot(MODEL2)
```

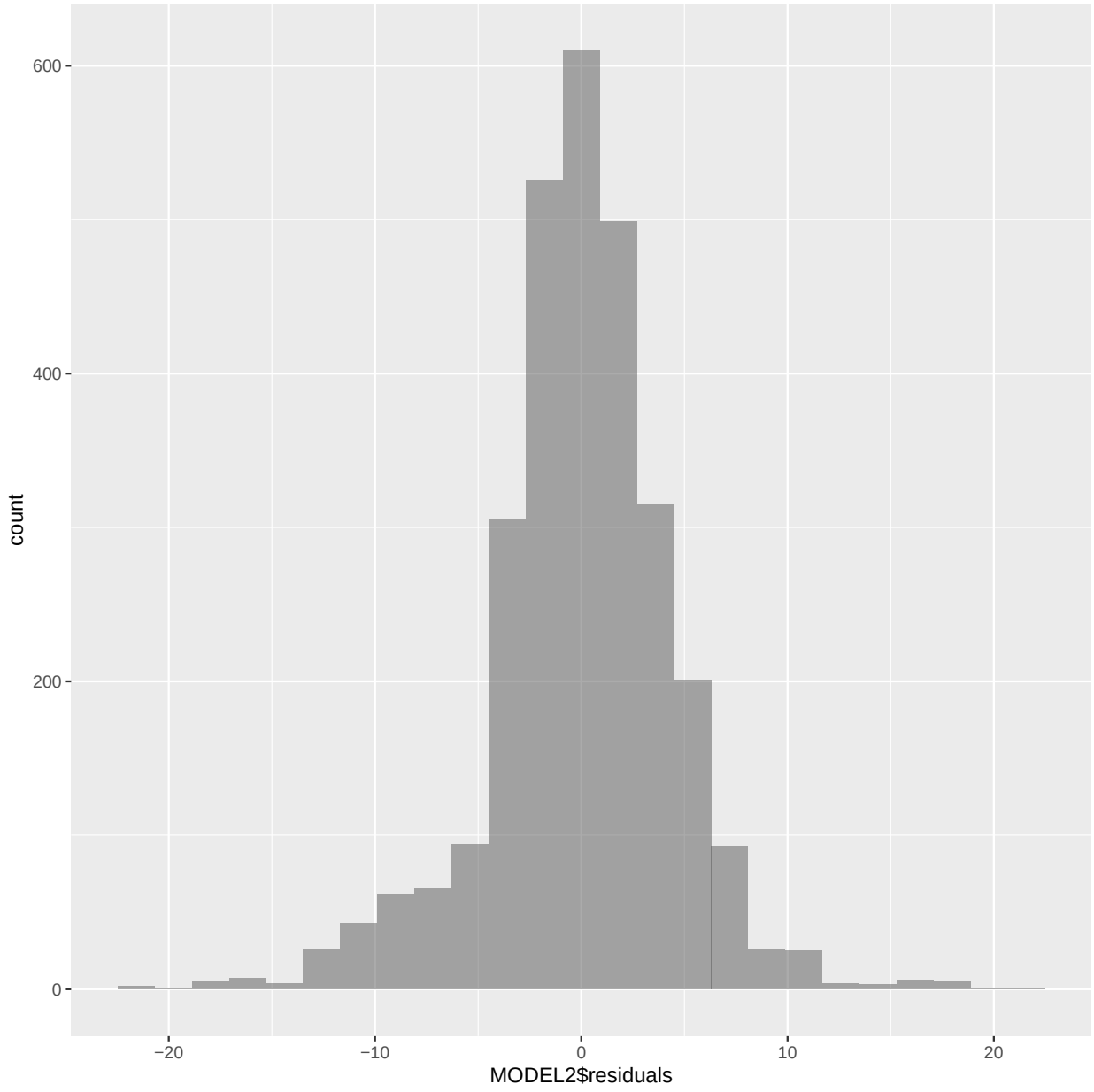

Residuals vs Fitted
Residuals
Fitted values
lm(Life.expectancy ~ Status + Adult.Mortality + Hepatitis.B + BMI + GDP + I ...
76
2308
2310
20
10
0
−10
−20
40
50
60
70
80
90


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Life.expectancy ~ Status + Adult.Mortality + Hepatitis.B + BMI + GDP + I ...
76
2308
2310
−4
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
2310
2308
76
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
40
50
60
70
80
90
Fitted values
lm(Life.expectancy ~ Status + Adult.Mortality + Hepatitis.B + BMI + GDP + I ...




```
gf_histogram(~MODEL2$residuals)
```




```
mean(MODEL2$residual^2)
```

```
## [1] 19.51552
```

```
summ(MODEL2, confint = TRUE, digits = 3)
```

```
## MODEL INFO:
## Observations: 2928 (10 missing obs. deleted)
## Dependent Variable: Life.expectancy
## Type: OLS linear regression
##
## MODEL FIT:
## F(8,2919) = 1330.409, p = 0.000
```

```
## R² = 0.785
## Adj. R² = 0.784
##
## Standard errors: OLS
## ------------------------------------------------------------------------------------
##                                        Est.     2.5%    97.5%    t val.        p
## ------------------------------------ -------- -------- -------- --------- -------
## (Intercept)                          56.985   55.758   58.212    91.053   0.000
## StatusDeveloping                     -1.859   -2.372   -1.345    -7.099   0.000
## Adult.Mortality                      -0.021   -0.023   -0.020   -24.792   0.000
## Hepatitis.B                           0.012    0.005    0.019     3.352   0.001
## BMI                                   0.062    0.053    0.072    12.628   0.000
## GDP                                   0.000    0.000    0.000     5.960   0.000
## Income.composition.of.resources       6.837    5.495    8.180     9.985   0.000
## HIV.AIDS                             -0.484   -0.521   -0.447   -25.528   0.000
## Schooling                             0.844    0.756    0.931    18.898   0.000
## ------------------------------------------------------------------------------------
```