Regression Analysis: Term Project

Libby Amir & Carmel Raviv

2023-12-08

Introduction:

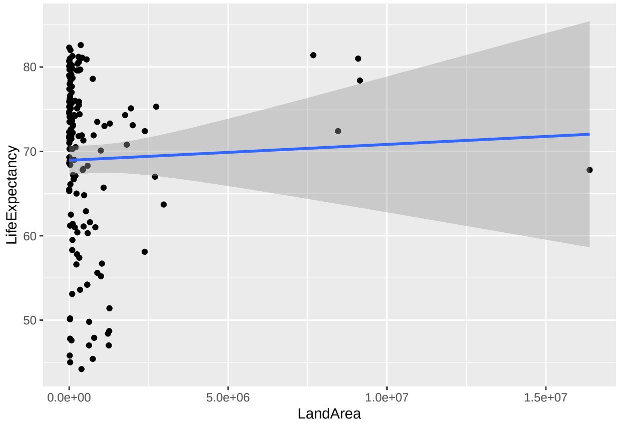
Our countries data gives us lots of information including the name of countries, land area in km^2, population in millions, % of population living in rural areas, % of government expenditures towards health care, % of population with internet access, birthrate per 1000 people, % of population at least 65 years old, average life expectancy in years, CO2 emissions in metric tons per capita, Gross Domestic Product per capita, and cell phone subscriptions per 100 people.

Using this data, we would like to understand which variables best contribute to the prediction of longevity. Some of the variables may have little to no effect on life expectancy, and should therefore be removed as to avoid confusing our regression model. By performing a residual analysis, transforming the variables when necessary, assessing the potential for multicollinearity, and executing a model selection algorithm, we can optimize a regression model for predicting a country's average life expectancy.

Simple Linear Regression Analysis:

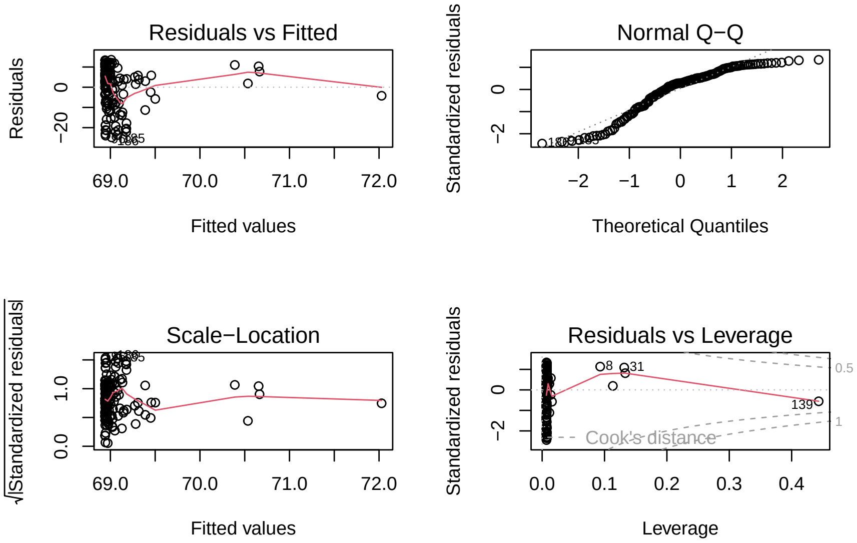
We can start the analysis by creating individual linear regression models of life expectancy against each of the other quantitative variables. We will not include qualitative/categorical variables in this because they cannot be regressed upon. Thus, we will not create regression models for the country names or their codes. To begin the regression analysis, we always create scatter plots for each of the predictor variables to see if the relationship seems linear. Our data includes some obscure, non-linear scatter plots, such as the example below for land area.

Example of Scatter Plots using Land Area (repeated for all variables):



To get more insight on the variables and assumptions, we need to obtain more plots for each of the linear models. These plots include the residual versus fitted values plot, the normal qq-plot, the scale-location plot, and the residuals versus leverage plot. From the shapes of these plots we can get a better idea of which variables are relevant to life expectancy, as well as which variables may need to undergo a transformation in order to be a valid predictor. For example, the residual plot for the land area predictor is highly clustered with a few extreme outliers. This is not ideal because a randomly scattered residual plot usually indicates linearity. Therefore, we can assume based on the land area residual plot that it most likely does not have a linear relationship with life expectancy. Another key element of our residual analysis is the normal qq-plots, which allow us to assess the normality of residuals. The ideal is to see a nearly straight line of dots translating to approximately normally distributed residuals. According to our qq-plots, we can identify many of the variables would need a box-cox transformation, including land area as shown in the example below. However, we will hold off on doing these transformations since we will proceed with multiple regression and model selection. We might eliminate these variables altogether so it might be ineffective to transform them individually.

Example of Residual Plots using Land Area (repeated for all variables):



Assessing Multicollinearity Using VIF:

Before proceeding with model selection, we want to check for multicollinearity between the variables. To do this we can use the vif() function which returns the VIF values for each of the predictor variables in regression model. Our model results in VIF values between 1 and 6, suggesting that there may be little to moderate multicollinearity between the variables. The variable with the highest potential of multicollinearity is Internet, so we can ponder what might be the cause for this. Perhaps the percentage of the population with access to internet is related to how high the GDP of that country is, or how many people have cellphone plans. This is important to acknowledge, but since there are no extreme VIF values, for the sake of simplicity we will go through with model selection despite this.

VIF Values:

```
##   LandArea Population      Rural     Health   Internet  BirthRate ElderlyPop
##   1.239074   1.177570   2.383121   1.310762   5.693166   4.179054   3.826519
##        CO2        GDP       Cell
##   2.042874   3.488824   2.177292
```

Model Selection Process:

For model selection, different methods will result in different models and levels of fit. In this case, we wanted to show a step-by-step process, so we chose backwards selection. Backwards Selection starts with all of the variables and one-by-one removes those with the least effect on Life Expectancy.

We started with model1 that regressed Life Expectancy against the rest of the possible predictors. Then we summarized model1 and obtained its anova table to analyze the effect of each predictor on the model. From the summary we get the p-values for a t-test testing the null hypothesis that $B_1 = 0$ implying the slope is zero. In parallel, from the anova table we get the p-values for the f-test testing whether or not each variable has an effect on the response variable. Both of these p-values reflect the probabilities that the variables have no relationship with life expectancy, so in this case we want to remove the variables with the highest p-values. For model1, the variable we should omit is Population because its p-value is 0.67625 which is the highest of all the predictors. Subsequently, in the anova table we can reaffirm this decision because the p-value of Population is 0.77, meaning that the probability of Population having NO effect on Life Expectancy is very high.

Output for Summary of Model1 with P-values and Adjusted R^2:

```
##
## Call:
## lm(formula = LifeExpectancy ~ LandArea + Population + Rural +
##     Health + Internet + BirthRate + ElderlyPop + CO2 + GDP +
##     Cell, data = countries_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9278  -2.4981   0.3888   2.9936  10.5689
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.490e+01  3.484e+00  24.365  < 2e-16 ***
## LandArea    -2.761e-07  2.234e-07  -1.236  0.21867
## Population   1.772e-03  4.235e-03   0.418  0.67625
## Rural       -6.703e-02  2.625e-02  -2.554  0.01175 *
## Health       2.553e-01  9.985e-02   2.556  0.01166 *
## Internet     5.386e-02  3.435e-02   1.568  0.11919
## BirthRate   -7.076e-01  7.771e-02  -9.106 9.14e-16 ***
## ElderlyPop  -4.440e-01  1.514e-01  -2.933  0.00393 **
## CO2         -4.156e-02  8.655e-02  -0.480  0.63186
## GDP          3.862e-05  3.999e-05   0.966  0.33586
## Cell         1.724e-02  1.301e-02   1.325  0.18738
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.756 on 137 degrees of freedom
## Multiple R-squared:  0.7955, Adjusted R-squared:  0.7806
## F-statistic:  53.3 on 10 and 137 DF,  p-value: < 2.2e-16
```

It is also important that we note the initial adjusted R^2 from Model1 is 0.7806 which signifies a moderately-strong, positive relationship between the predictors and life expectancy. As we remove variables, it is important that the magnitude of the R^2 increases, because when it decreases then we have removed too many variables and the model is weaker. Thus, we continue this process until we see that our adjusted R^2 goes down instead of up.

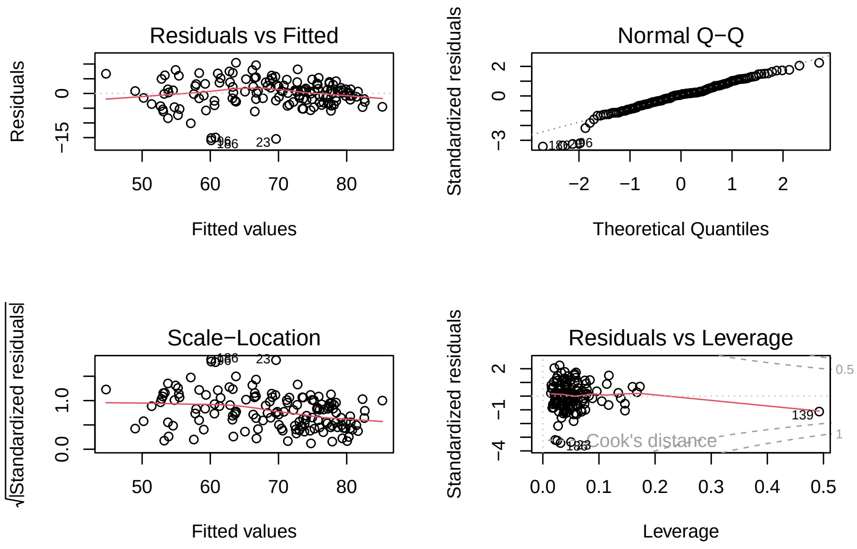
Next we make model2 that excludes Population, and get the summary statistics and anova table. Our model2 has a new adjusted R^2 is 0.7819, so we can see that removing the population variable has made our data more linear and a better fit for the model. The highest p-value in our new model is from the CO2 variable, at a 0.59943 probability of having a slope of zero. Since this suggests that it probably does not have an effect on Life Expectancy, we can now remove the CO2 variable.

Model3 has all predictors except Population and CO2, with a new adjusted R^2 of 0.783. This is a sign that our model is still improving, so we can try again to remove another variable. The highest p-value in our new model is GDP, a 0.38073 probability of having a slope of zero, so we omit it from model4. Model4 now has all predictors except for Population, CO2, and GDP, resulting in a new adjusted R^2 of 0.7834. Repeating the same process, the new highest p-value in our model is from the Cell variable, with a 0.21513 probability of having a slope of zero, suggesting that it is not a good predictor for Life Expectancy. Now model5 has all predictors except for Population, CO2, GDP, and Cell, giving us a new adjusted R^2 is 0.7825. This has a smaller magnitude than our previous adjusted R^2 in model4, so our model did not improve by omitting the Cell variable. This means our best model was model4 and we do not need to proceed with our selection.

Final Round of Residual Analysis on Selected Model:

Our optimal model proves that the predictors of Life Expectancy are LandArea, Rural, Health, Internet, BirthRate, ElderlyPop, and Cell. With this model we can now perform another round of residual analysis to check it would benefit from any transformations. Similar to with the simple linear models, we obtain the various plots and analyze if the regression assumptions are met. We see that our residual plots are pretty decent, but it is unclear if there is normality in the residuals, so we are going to consider a box-cox transformation.

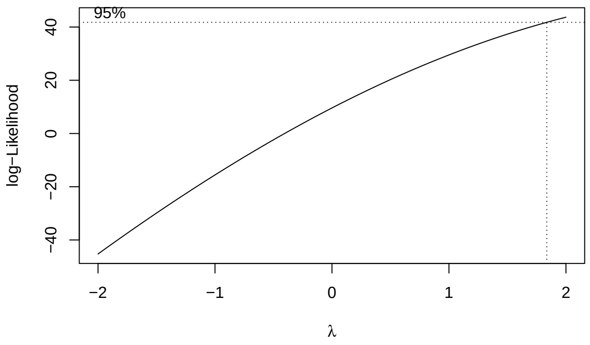
Plots for Residual Analysis:



Boxcox Transformation:

By creating a boxcox plot of our selected model, we can see that our lambda is around 1.8, so we can perform a box-cox transformation using this value.

Boxcox Plot:



After transforming the response variable, Life Expectancy, we run the summary analysis to see if our model is better fitted with the transformed data. After executing this, we see that the transformation has slightly improved our adjusted R^2, which is now 0.7992. Now that we have had an effective transformation, we have reached our final model.

Conclusion:

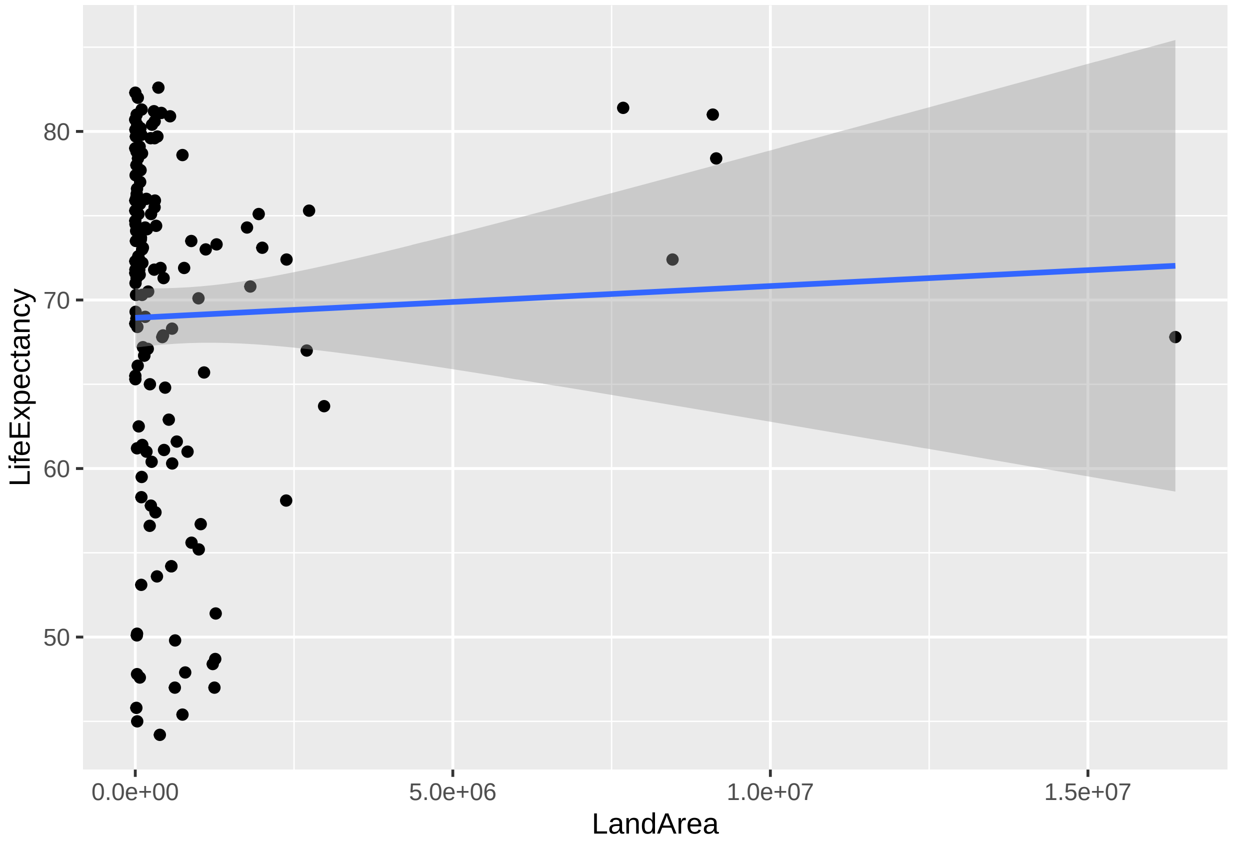
In conclusion, we have found that longevity is most impacted by a country's land area, % of rural, average health, internet access, birth rate, % of elderly population, and % of cellphone plans. These various factors have ties to average life expectancy, and can allow us to somewhat accurately predict what a country's life expectancy is without actually knowing.

Appendix

```
# load the necessary packages
library(tidyverse)
library(dplyr)
library(ggplot2)
library(MASS)
# obtain unique subset
countries_df <- read.csv("countries.csv")
n <- nrow(countries_df)
set.seed(920998496)
subset_id <- sample(n, 0.8*n)
countries_df <- countries_df[subset_id, ]
# create the regression models
land_lm <- lm(LifeExpectancy ~ LandArea, data = countries_df)
pop_lm <- lm(LifeExpectancy ~ Population, data = countries_df)
rural_lm <- lm(LifeExpectancy ~ Rural, data = countries_df)
health_lm <- lm(LifeExpectancy ~ Health, data = countries_df)
internet_lm <- lm(LifeExpectancy ~ Internet, data = countries_df)
birth_lm <- lm(LifeExpectancy ~ BirthRate, data = countries_df)
elderly_lm <- lm(LifeExpectancy ~ ElderlyPop, data = countries_df)
co2_lm <- lm(LifeExpectancy ~ CO2, data = countries_df)
gdp_lm <- lm(LifeExpectancy ~ GDP, data = countries_df)
cell_lm <- lm(LifeExpectancy ~ Cell, data = countries_df)
# scatter plots of each of the linear models
```

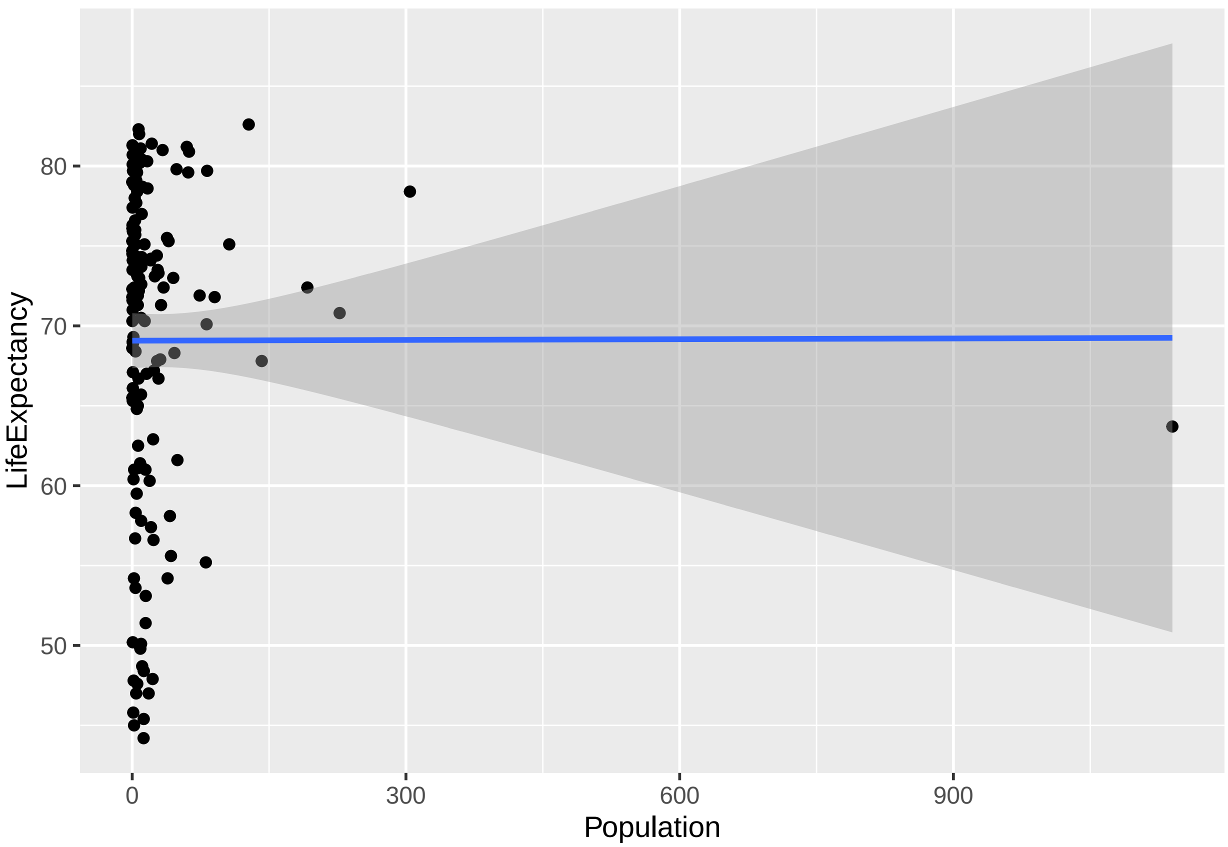
```
ggplot(land_lm, aes(x = LandArea, y = LifeExpectancy)) +  # referenced in the report as an ie.
  geom_point() +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



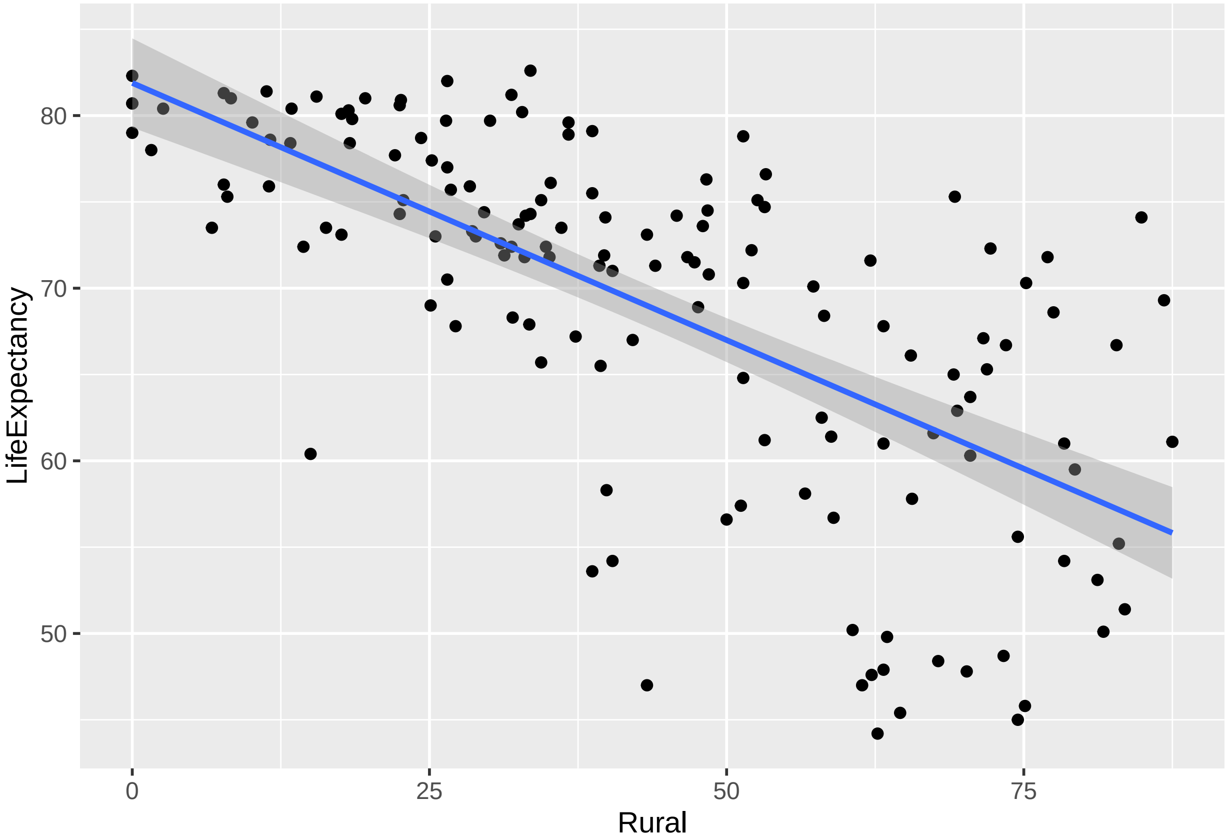
```
ggplot(pop_lm, aes(x = Population, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



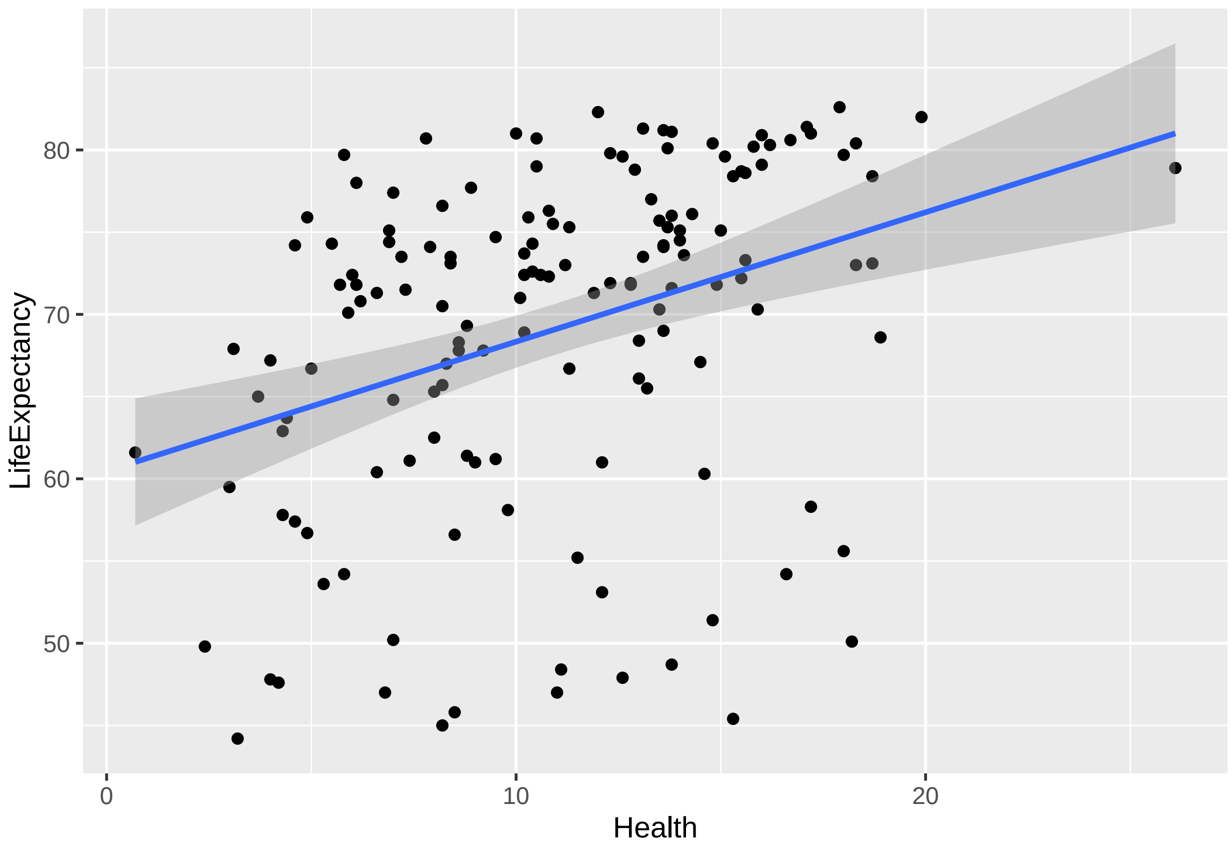
```
ggplot(rural_lm, aes(x = Rural, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



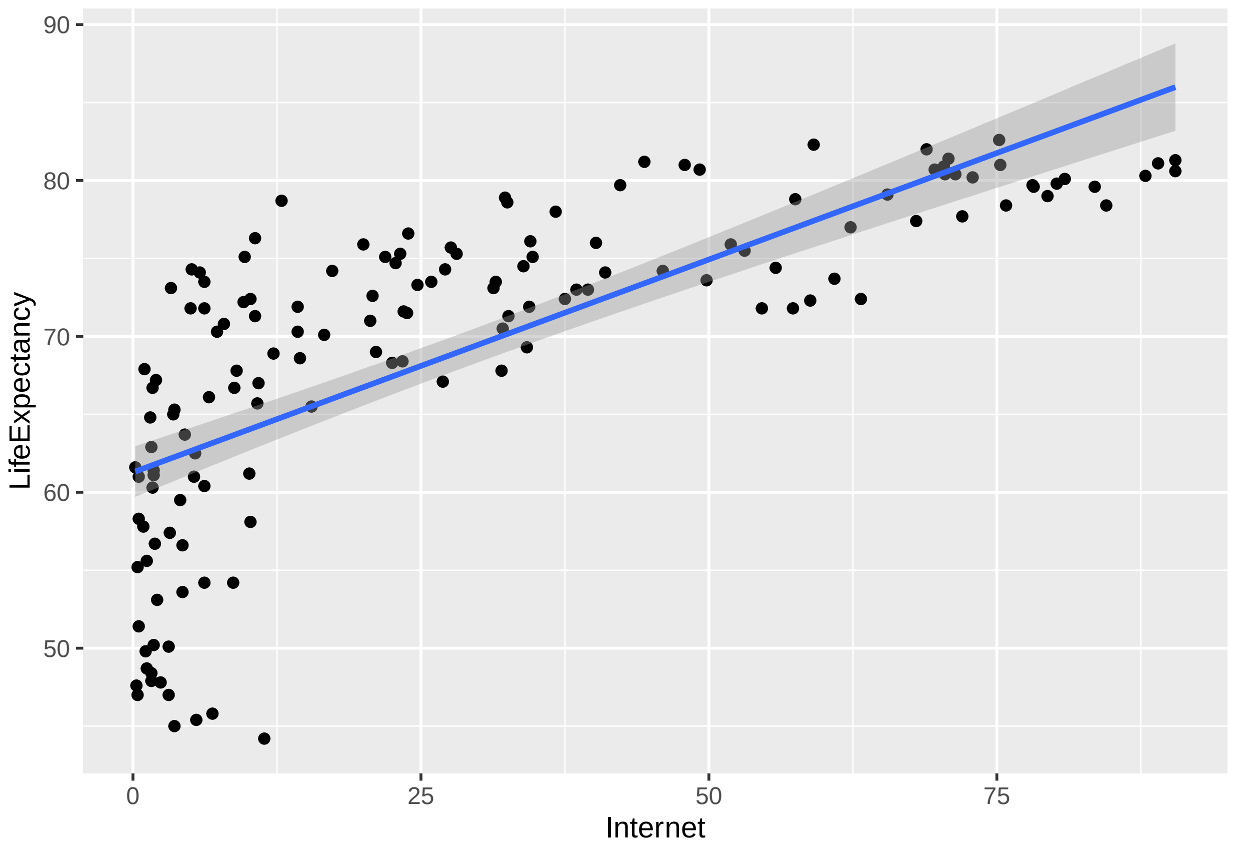
```
ggplot(health_lm, aes(x = Health, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



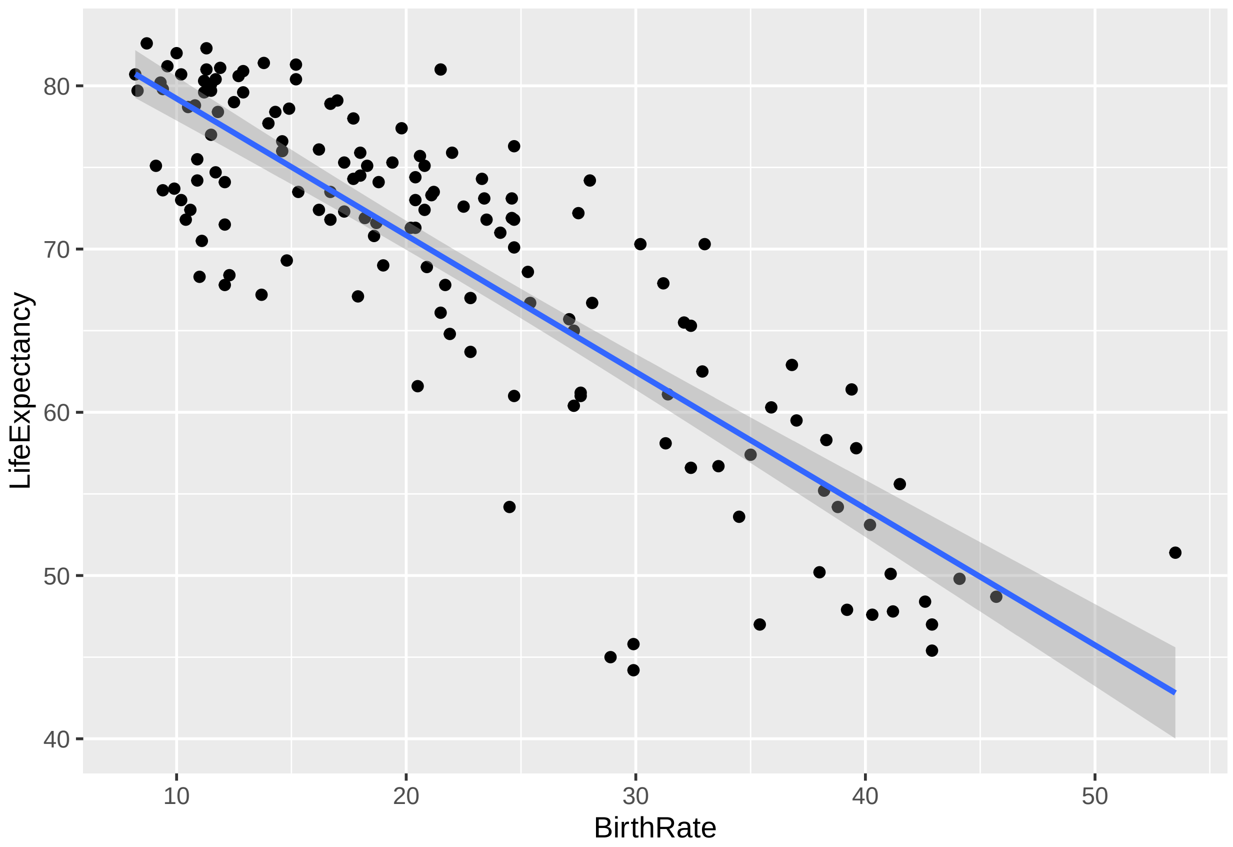
```
ggplot(internet_lm, aes(x = Internet, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



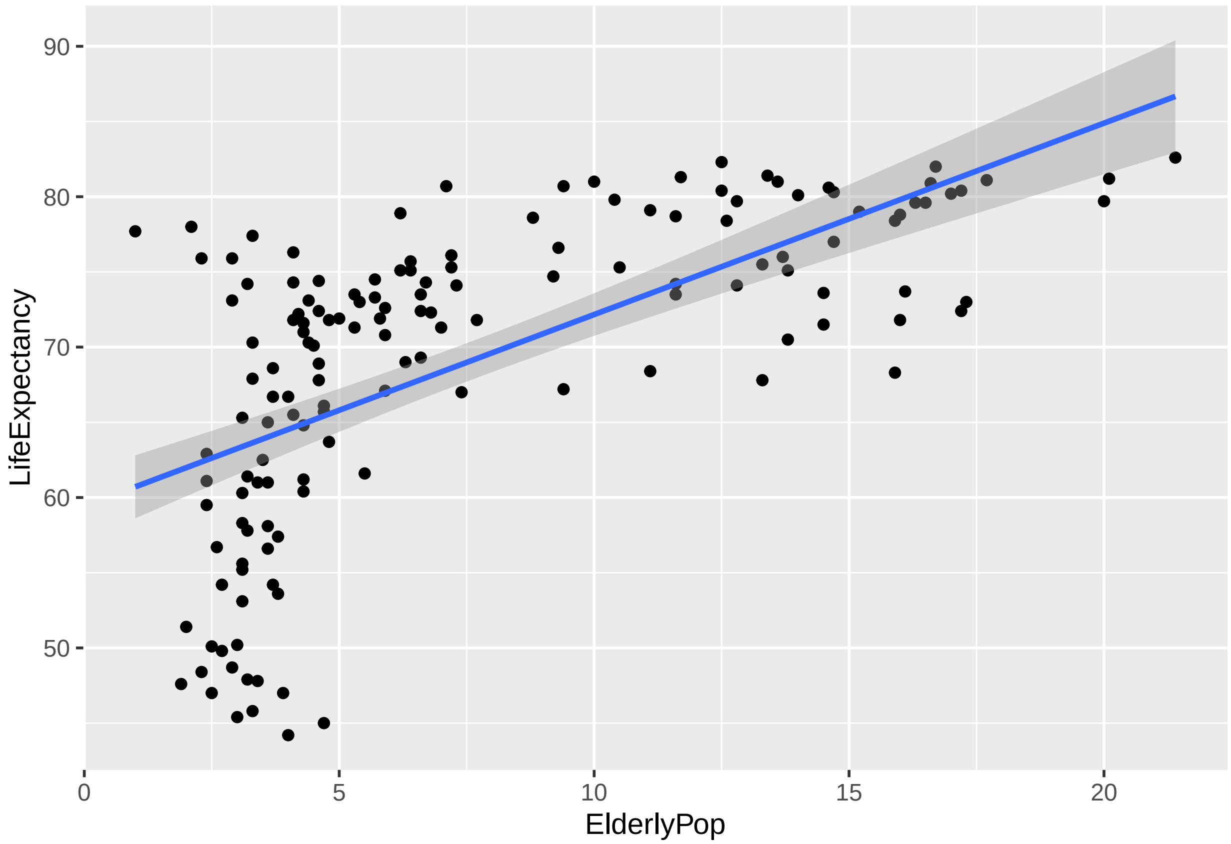
```
ggplot(birth_lm, aes(x = BirthRate, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



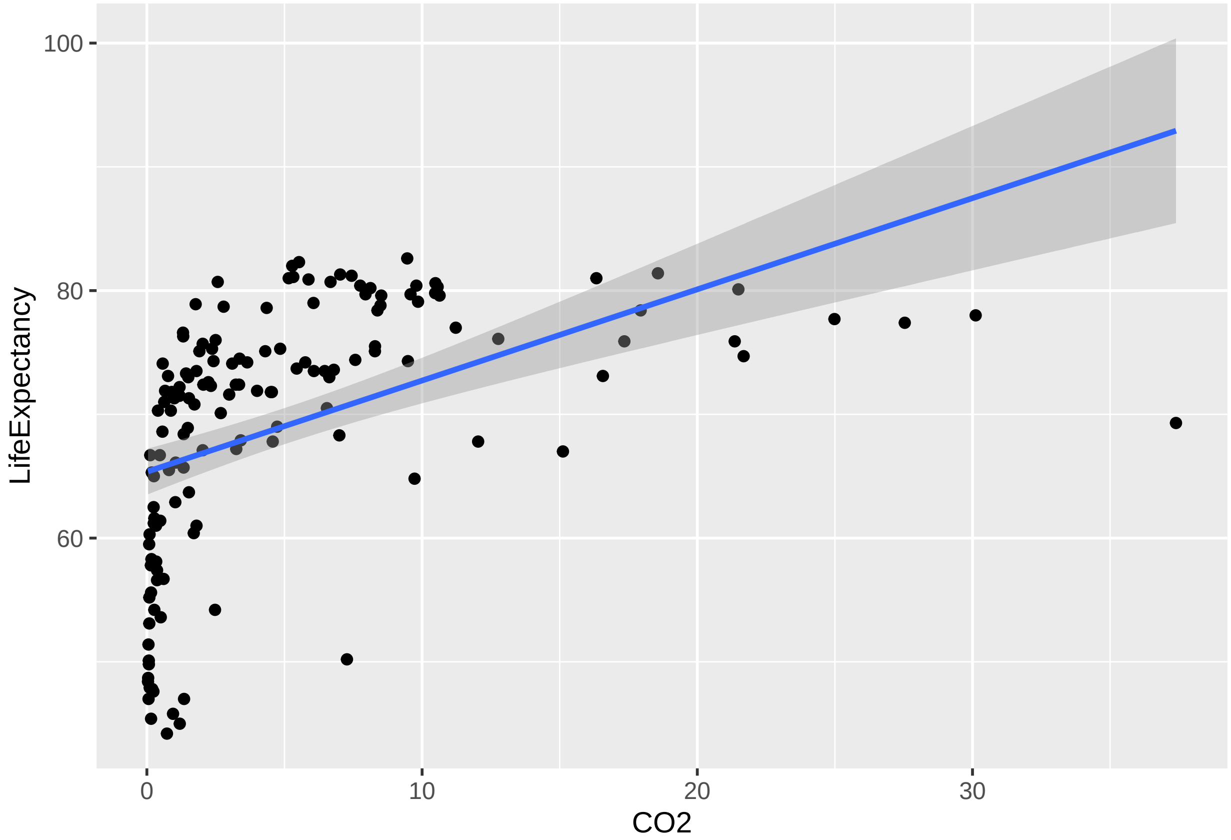
```
ggplot(elderly_lm, aes(x = ElderlyPop, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



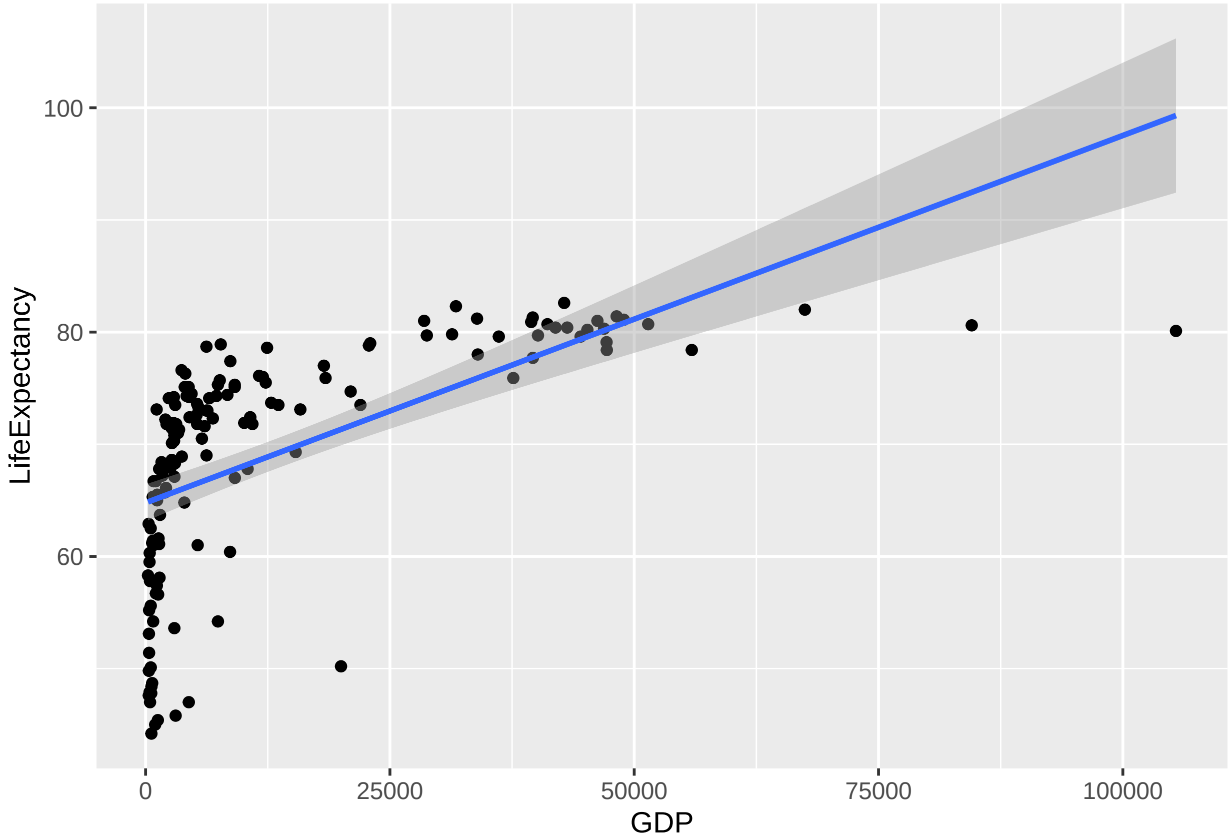
```
ggplot(co2_lm, aes(x = CO2, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



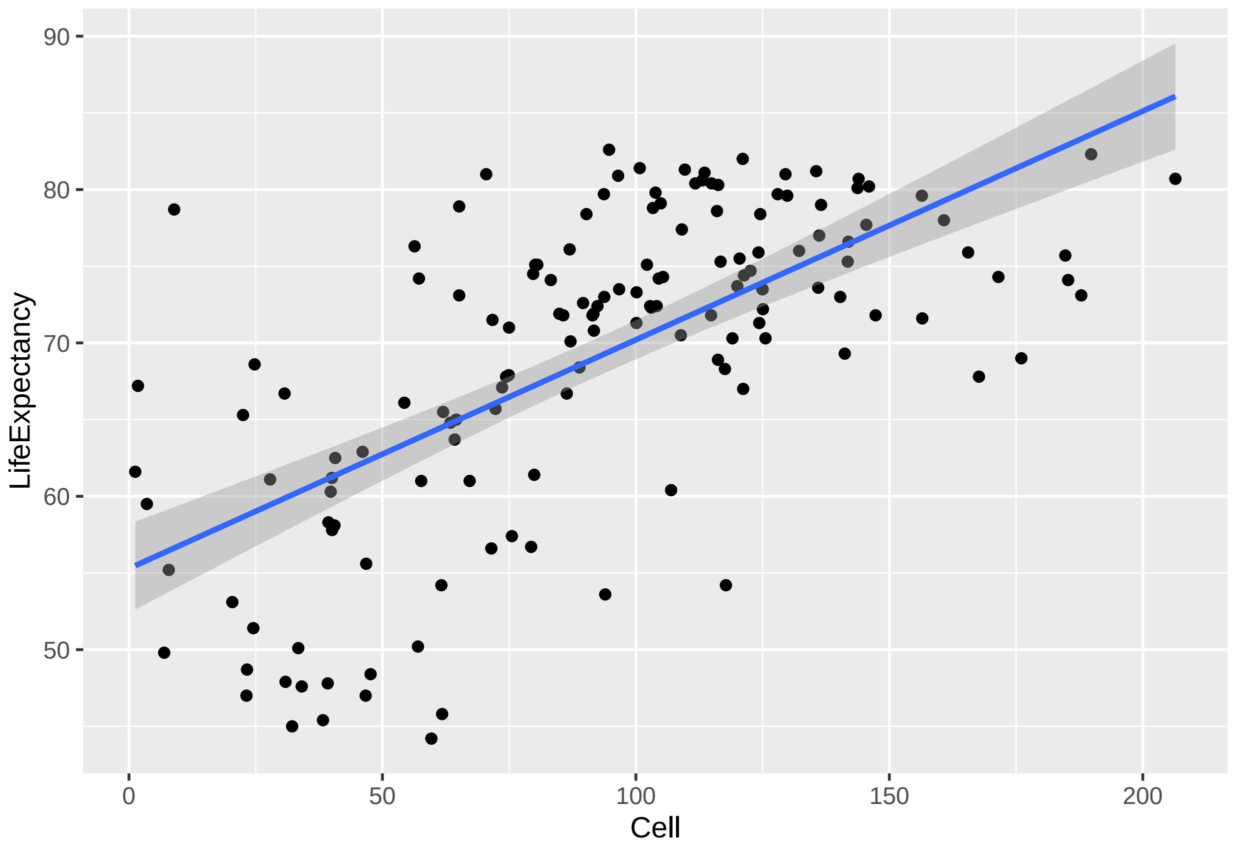
```
ggplot(gdp_lm, aes(x = GDP, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

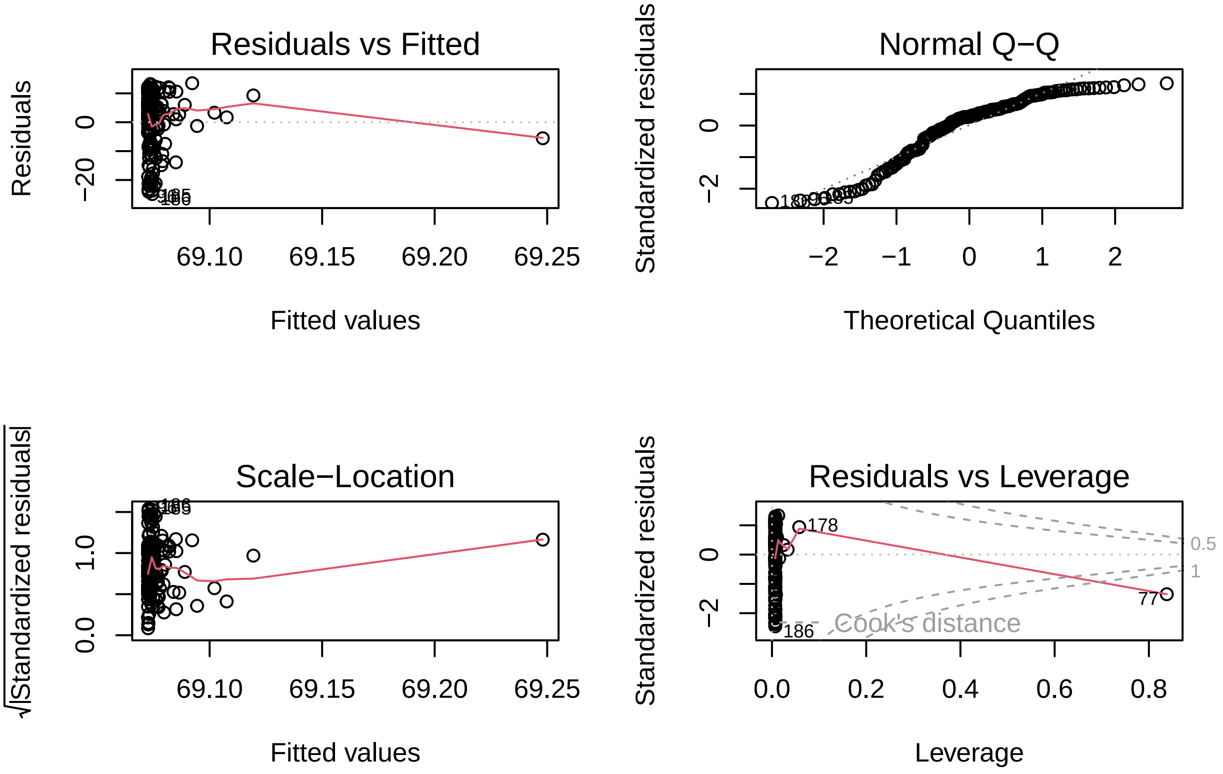


```
ggplot(cell_lm, aes(x = Cell, y = LifeExpectancy)) +
  geom_point() +
  stat_smooth(method = "lm")
```

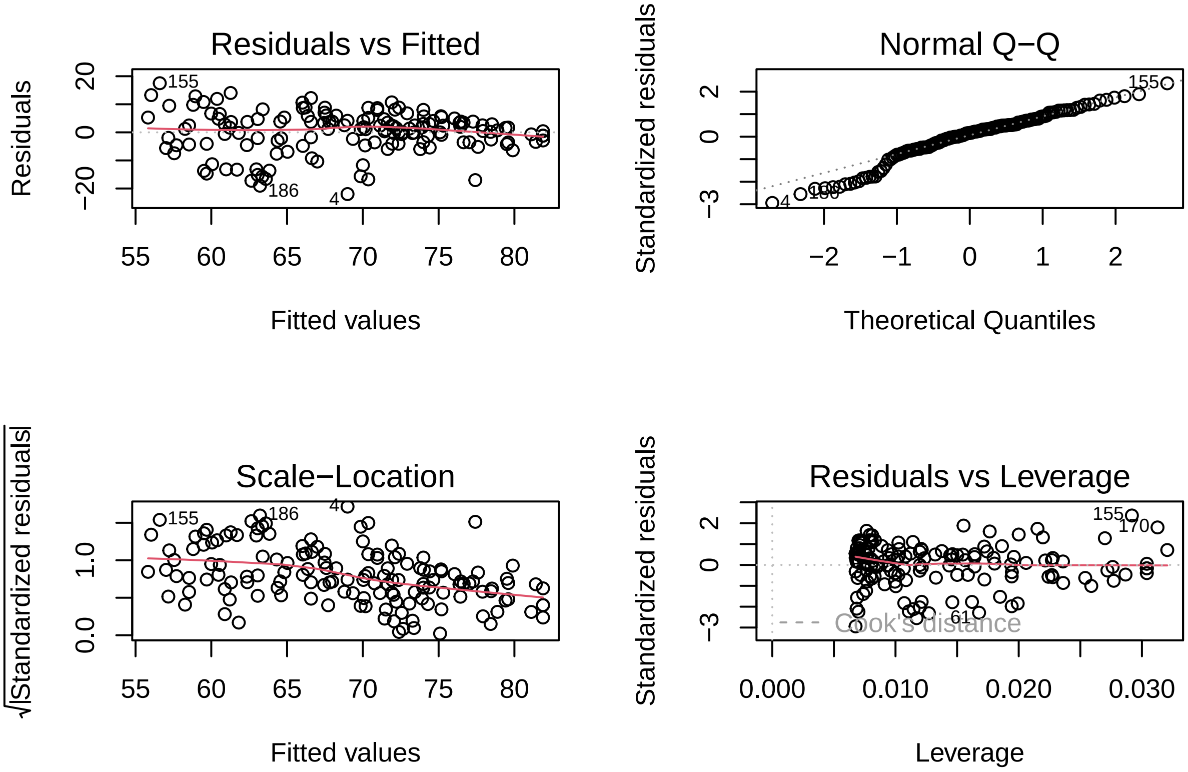
```
## `geom_smooth()` using formula 'y ~ x'
```



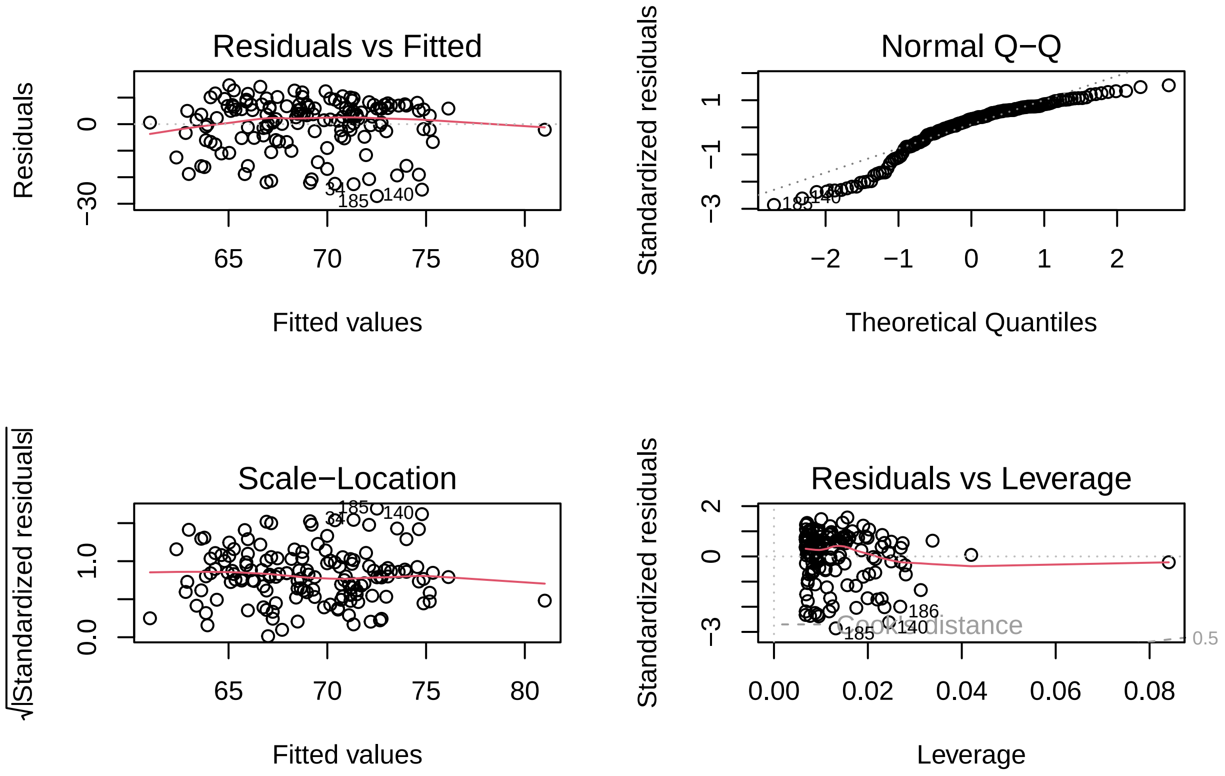
```
# plots of each of the linear models
par(mfrow = c(2, 2))
plot(pop_lm)
```



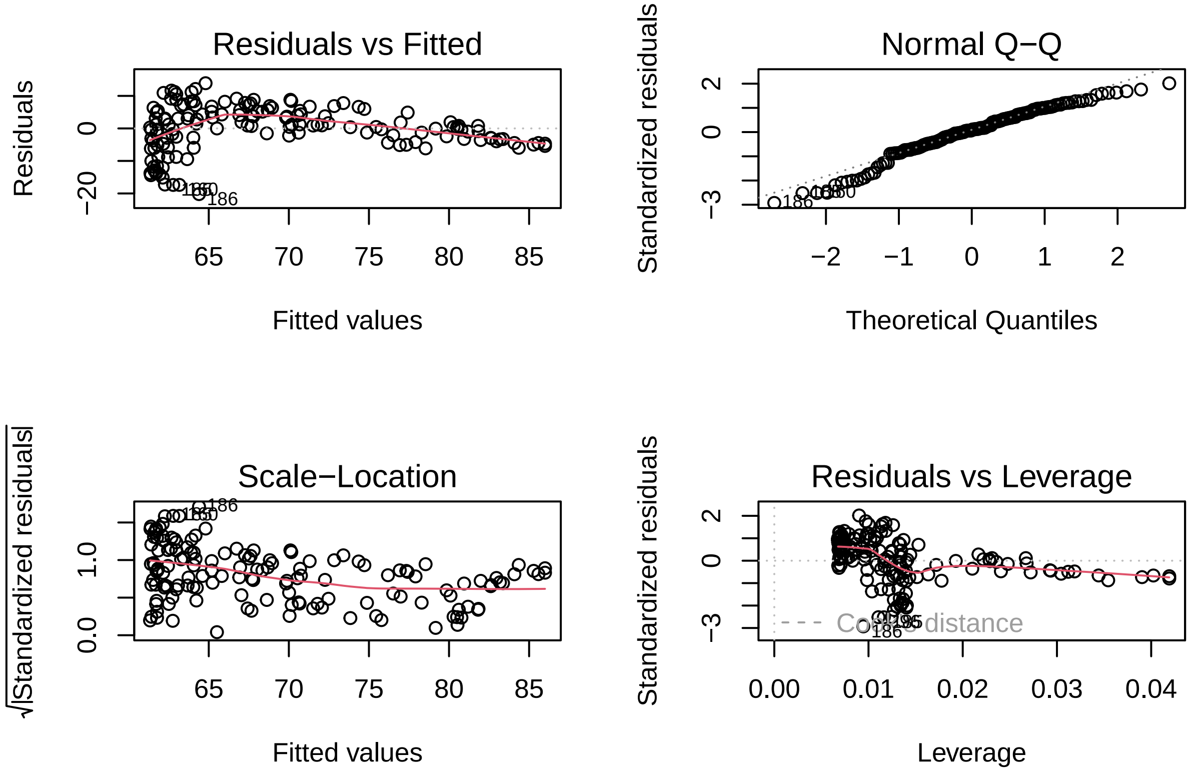
```
plot(rural_lm)
```



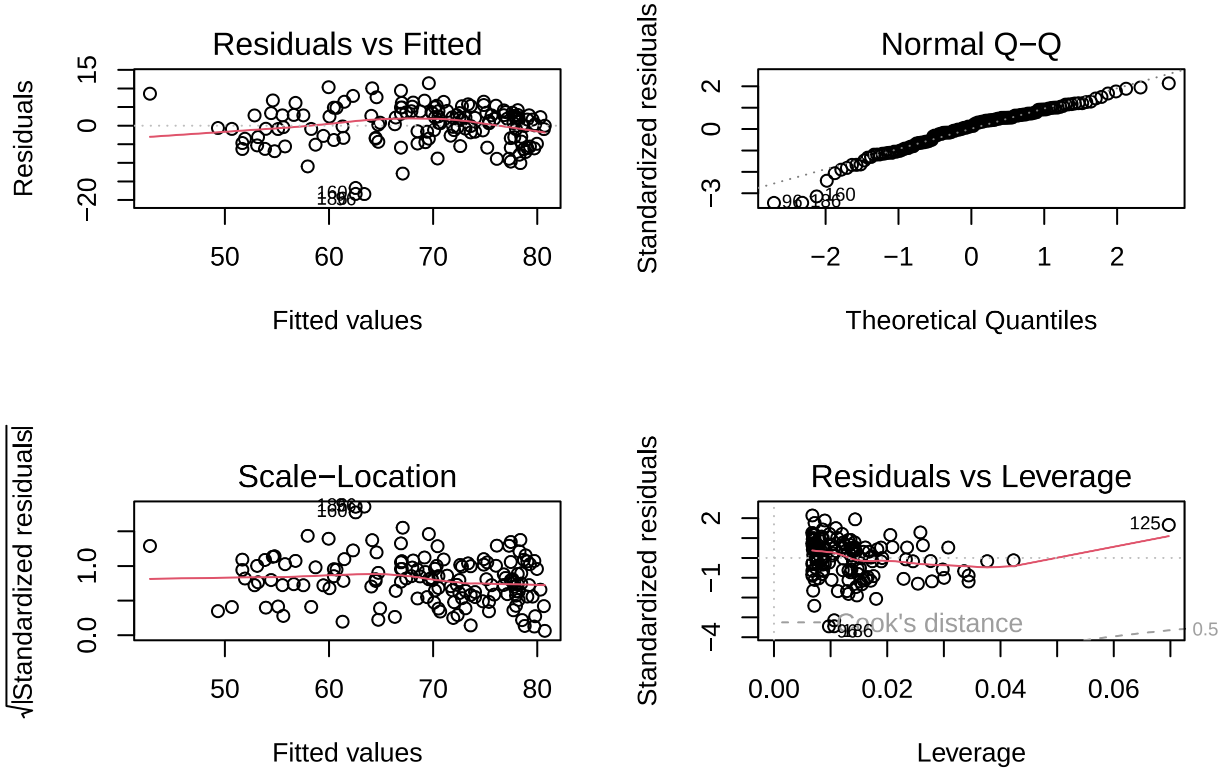
```
plot(health_lm)
```



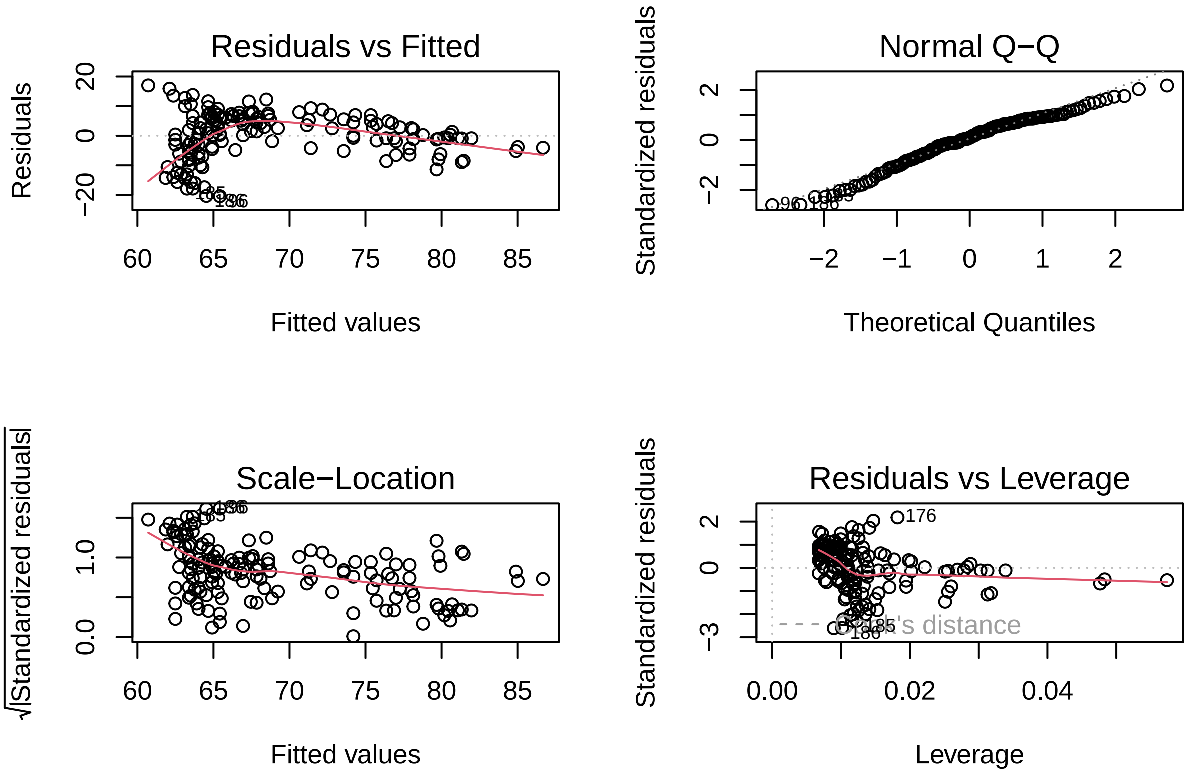
```
plot(internet_lm)
```



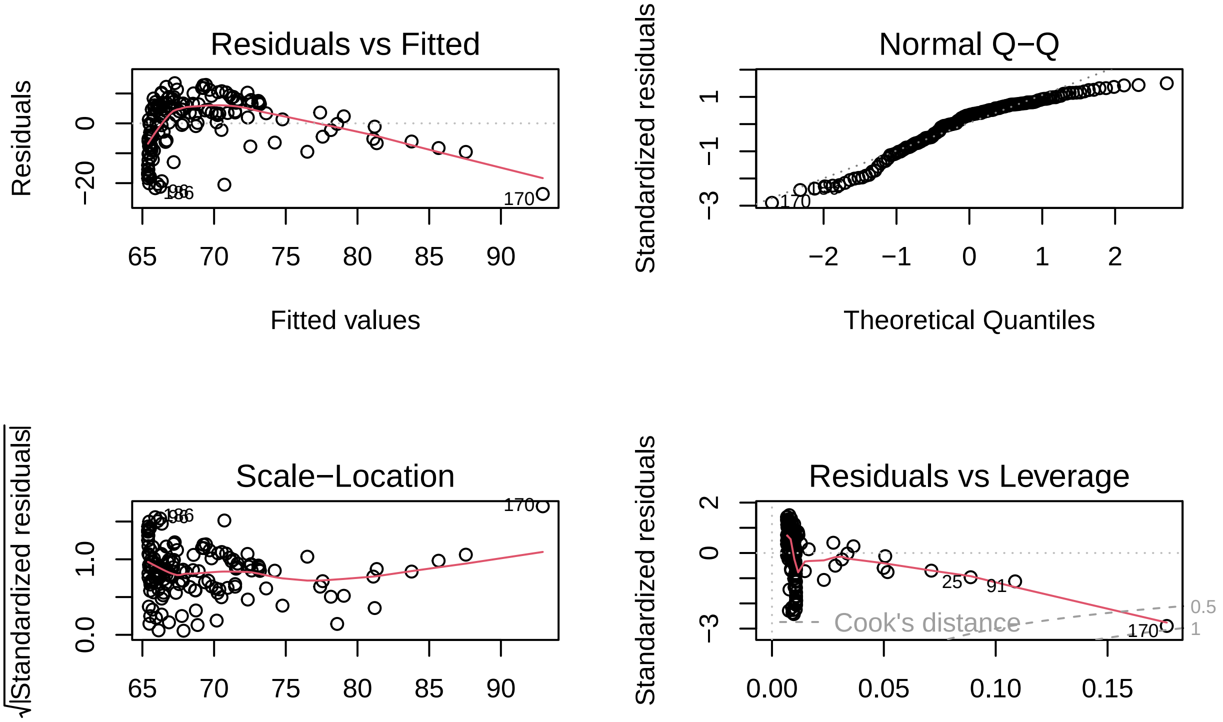
```
plot(birth_lm)
```



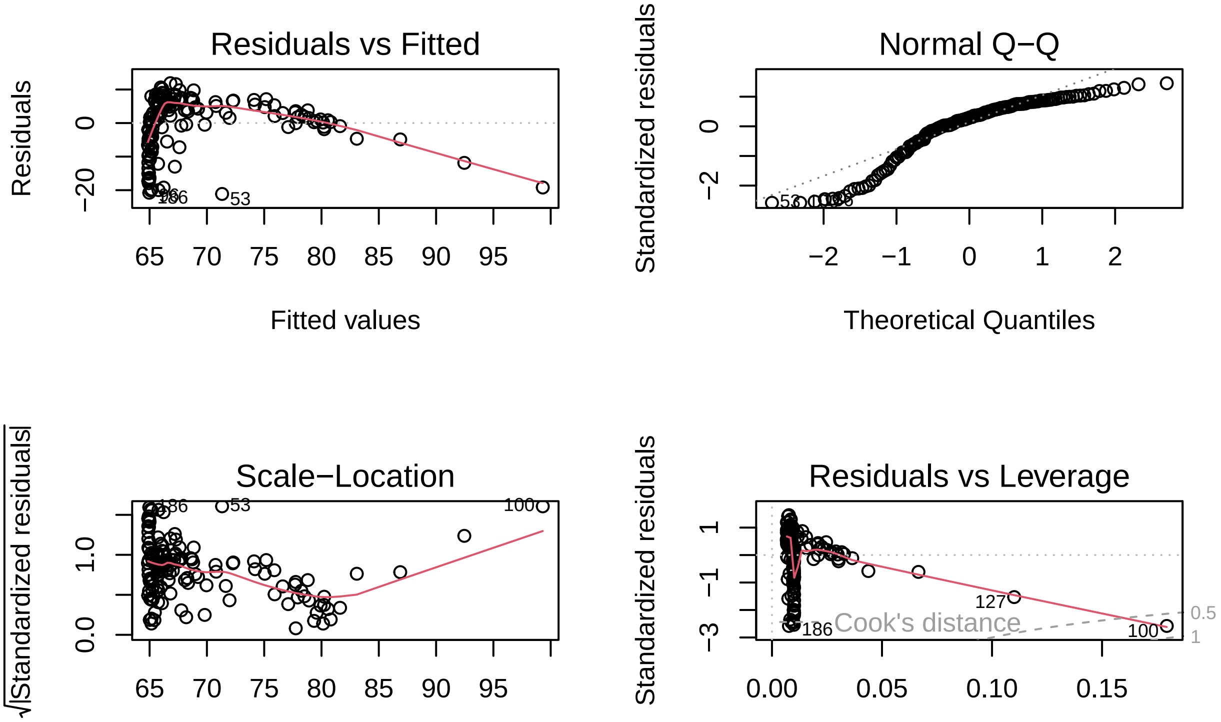
```
plot(elderly_lm)
```


Residuals vs Fitted
Residuals
Fitted values
20
0
−20
60
65
70
75
80
85
185
186
Normal Q−Q
Standardized residuals
Theoretical Quantiles
2
0
−2
−1
1
96
186
185
Scale−Location
√|Standardized residuals|
Fitted values
1.0
0.0
186
Residuals vs Leverage
Standardized residuals
Leverage
0.00
0.02
0.04
−3
176
186
185
Cook's distance

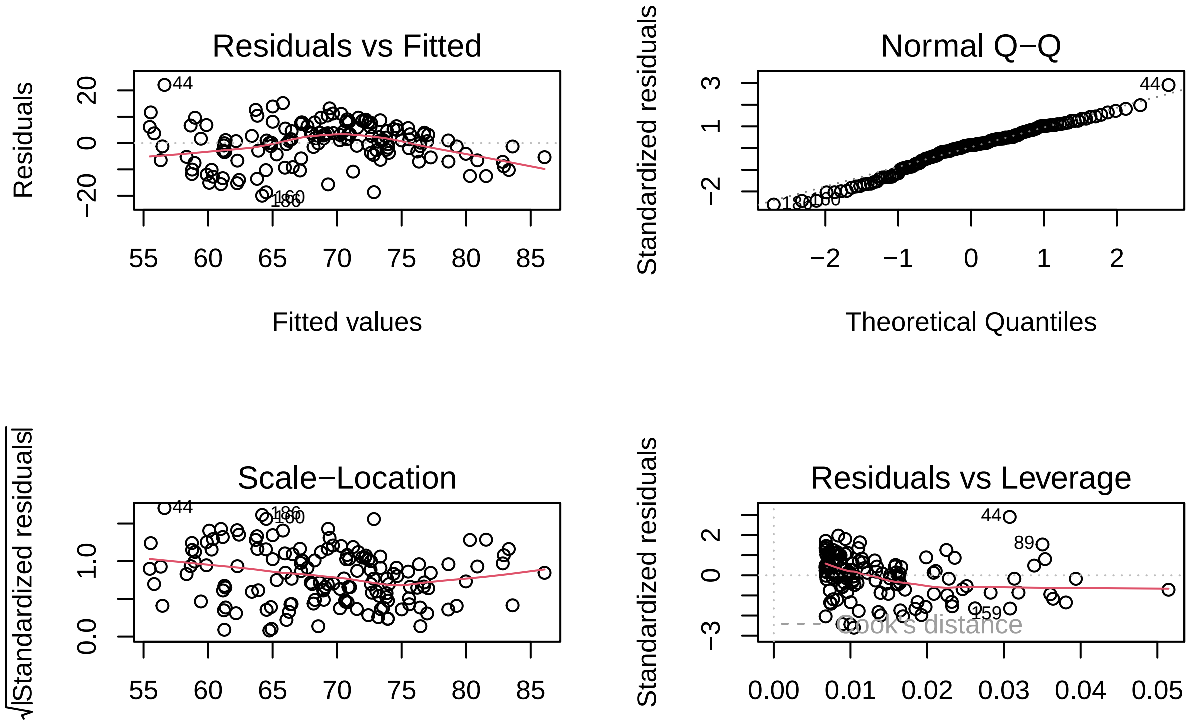
```
plot(co2_lm)
```


Residuals vs Fitted
Residuals
0
−20
65
70
75
80
85
90
Fitted values
186
170
Normal Q−Q
Standardized residuals
1
−1
−3
−2
−1
0
1
2
Theoretical Quantiles
170
Scale−Location
√|Standardized residuals|
1.0
0.0
65
70
75
80
85
90
Fitted values
186
170
Residuals vs Leverage
Standardized residuals
2
0
−3
0.00
0.05
0.10
0.15
Leverage
Cook's distance
25
91
170
0.5
1

```
plot(gdp_lm)
```


Residuals vs Fitted
Residuals
0
−20
65 70 75 80 85 90 95
Fitted values
186
53
Normal Q−Q
Standardized residuals
0
−2
−2 −1 0 1 2
Theoretical Quantiles
53
Scale−Location
√|Standardized residuals|
0.0
1.0
65 70 75 80 85 90 95
Fitted values
186
53
100
Residuals vs Leverage
Standardized residuals
1
−1
−3
0.00 0.05 0.10 0.15
Leverage
127
186
100
Cook's distance
0.5
1

```
plot(cell_lm)
```



```
#Obtain summary and anova table for regression model with all variables
summary(model1) # referred to in
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ LandArea + Population + Rural +
##     Health + Internet + BirthRate + ElderlyPop + CO2 + GDP +
##     Cell, data = countries_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9278  -2.4981   0.3888   2.9936  10.5689
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.490e+01  3.484e+00  24.365  < 2e-16 ***
## LandArea    -2.761e-07  2.234e-07  -1.236  0.21867
## Population   1.772e-03  4.235e-03   0.418  0.67625
## Rural       -6.703e-02  2.625e-02  -2.554  0.01175 *
## Health       2.553e-01  9.985e-02   2.556  0.01166 *
## Internet     5.386e-02  3.435e-02   1.568  0.11919
## BirthRate   -7.076e-01  7.771e-02  -9.106 9.14e-16 ***
## ElderlyPop  -4.440e-01  1.514e-01  -2.933  0.00393 **
## CO2         -4.156e-02  8.655e-02  -0.480  0.63186
## GDP          3.862e-05  3.999e-05   0.966  0.33586
## Cell         1.724e-02  1.301e-02   1.325  0.18738
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.756 on 137 degrees of freedom
## Multiple R-squared:  0.7955, Adjusted R-squared:  0.7806
## F-statistic:  53.3 on 10 and 137 DF,  p-value: < 2.2e-16
```

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: LifeExpectancy
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## LandArea      1   20.0    20.0   0.8825  0.349185
## Population    1    1.8     1.8   0.0800  0.777732
## Rural         1 7054.2  7054.2 311.9186 < 2.2e-16 ***
## Health        1  688.0   688.0  30.4227 1.675e-07 ***
## Internet      1 1562.7  1562.7  69.0970 8.233e-14 ***
## BirthRate     1 2442.6  2442.6 108.0064 < 2.2e-16 ***
## ElderlyPop    1  225.2   225.2   9.9590  0.001968 **
## CO2           1    0.0     0.0   0.0011  0.973753
## GDP           1   18.7    18.7   0.8285  0.364314
## Cell          1   39.7    39.7   1.7556  0.187381
## Residuals   137 3098.3    22.6
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#Running regression model with all variables except Population
model2 <- lm(LifeExpectancy~LandArea+Rural+Health+Internet+BirthRate+ElderlyPop+CO2+GDP+Cell,
             data = countries_df)
summary(model2)
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ LandArea + Rural + Health + Internet +
##     BirthRate + ElderlyPop + CO2 + GDP + Cell, data = countries_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9964  -2.5279   0.4568   2.9001  10.5441
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.509e+01  3.442e+00  24.721  < 2e-16 ***
## LandArea    -2.430e-07  2.083e-07  -1.166  0.24546
## Rural       -6.594e-02  2.604e-02  -2.532  0.01245 *
## Health       2.500e-01  9.876e-02   2.531  0.01248 *
## Internet     5.376e-02  3.425e-02   1.570  0.11873
## BirthRate   -7.122e-01  7.671e-02  -9.285  3.1e-16 ***
## ElderlyPop  -4.468e-01  1.508e-01  -2.963  0.00358 **
## CO2         -4.520e-02  8.586e-02  -0.526  0.59943
## GDP          3.969e-05  3.978e-05   0.998  0.32018
## Cell         1.692e-02  1.295e-02   1.306  0.19359
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.741 on 138 degrees of freedom
## Multiple R-squared:  0.7952, Adjusted R-squared:  0.7819
## F-statistic: 59.55 on 9 and 138 DF,  p-value: < 2.2e-16
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: LifeExpectancy
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## LandArea     1   20.0    20.0   0.8878  0.347732
## Rural        1 7009.6  7009.6 311.8095 < 2.2e-16 ***
## Health       1  648.1   648.1  28.8286 3.263e-07 ***
## Internet     1 1594.8  1594.8  70.9427 4.263e-14 ***
## BirthRate    1 2493.4  2493.4 110.9138 < 2.2e-16 ***
## ElderlyPop   1  225.1   225.1  10.0111  0.001914 **
## CO2          1    0.0     0.0   0.0000  0.995257
## GDP          1   19.8    19.8   0.8797  0.349936
## Cell         1   38.4    38.4   1.7067  0.193586
## Residuals  138 3102.3    22.5
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#Running regression model with all variables except Population & CO2
model3 <- lm(LifeExpectancy~LandArea+Rural+Health+Internet+BirthRate+ElderlyPop+GDP+Cell,
             data = countries_df)
summary(model3)
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ LandArea + Rural + Health + Internet +
##     BirthRate + ElderlyPop + GDP + Cell, data = countries_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8720  -2.6210   0.4067   2.8882  10.5638
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.473e+01  3.365e+00  25.182  < 2e-16 ***
## LandArea    -2.662e-07  2.030e-07  -1.311  0.19189
## Rural       -6.771e-02  2.575e-02  -2.629  0.00953 **
## Health       2.591e-01  9.698e-02   2.672  0.00844 **
## Internet     5.012e-02  3.345e-02   1.498  0.13636
## BirthRate   -7.020e-01  7.404e-02  -9.482  < 2e-16 ***
## ElderlyPop  -4.199e-01  1.415e-01  -2.968  0.00353 **
## GDP          3.314e-05  3.769e-05   0.879  0.38073
## Cell         1.569e-02  1.271e-02   1.235  0.21892
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 4.729 on 139 degrees of freedom
## Multiple R-squared:  0.7948, Adjusted R-squared:  0.783
## F-statistic: 67.31 on 8 and 139 DF,  p-value: < 2.2e-16
```

```
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: LifeExpectancy
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## LandArea      1   20.0    20.0   0.8924  0.346468
## Rural         1 7009.6  7009.6 313.4396 < 2.2e-16 ***
## Health        1  648.1   648.1  28.9793 3.032e-07 ***
## Internet      1 1594.8  1594.8  71.3136 3.622e-14 ***
## BirthRate     1 2493.4  2493.4 111.4937 < 2.2e-16 ***
## ElderlyPop    1  225.1   225.1  10.0634  0.001862 **
## GDP           1   17.8    17.8   0.7962  0.373786
## Cell          1   34.1    34.1   1.5252  0.218921
## Residuals   139 3108.5    22.4
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#Running regression model with all variables except Population, CO2, & GDP
model4 <- lm(LifeExpectancy~LandArea+Rural+Health+Internet+BirthRate+ElderlyPop+Cell,
             data = countries_df)
summary(model4)
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ LandArea + Rural + Health + Internet +
##     BirthRate + ElderlyPop + Cell, data = countries_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9261  -2.5178   0.3953   2.9600  10.4369
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.446e+01  3.347e+00  25.231  < 2e-16 ***
## LandArea    -2.628e-07  2.028e-07  -1.296  0.19719
## Rural       -7.125e-02  2.542e-02  -2.803  0.00577 **
## Health       2.650e-01  9.667e-02   2.742  0.00691 **
## Internet     6.810e-02  2.645e-02   2.574  0.01108 *
## BirthRate   -6.912e-01  7.295e-02  -9.475  < 2e-16 ***
## ElderlyPop  -4.152e-01  1.413e-01  -2.939  0.00385 **
## Cell         1.581e-02  1.270e-02   1.245  0.21513
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.725 on 140 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7834
## F-statistic: 76.94 on 7 and 140 DF,  p-value: < 2.2e-16
```

```
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: LifeExpectancy
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## LandArea      1   20.0    20.0   0.8938  0.346066
## Rural         1 7009.6  7009.6 313.9481 < 2.2e-16 ***
## Health        1  648.1   648.1  29.0263 2.946e-07 ***
## Internet      1 1594.8  1594.8  71.4293 3.352e-14 ***
## BirthRate     1 2493.4  2493.4 111.6745 < 2.2e-16 ***
## ElderlyPop    1  225.1   225.1  10.0797  0.001844 **
## Cell          1   34.6    34.6   1.5506  0.215126
## Residuals   140 3125.8    22.3
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#Running regression model with all variables except Population, CO2, GDP, & Cell
model5 <- lm(LifeExpectancy~LandArea+Rural+Health+Internet+BirthRate+ElderlyPop,
             data = countries_df)
summary(model5)
```

```
##
## Call:
## lm(formula = LifeExpectancy ~ LandArea + Rural + Health + Internet +
##     BirthRate + ElderlyPop, data = countries_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.047  -2.780   0.416   2.881   9.994
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.728e+01  2.466e+00  35.386  < 2e-16 ***
## LandArea    -2.727e-07  2.031e-07  -1.343 0.181508
## Rural       -8.190e-02  2.398e-02  -3.416 0.000832 ***
## Health       2.608e-01  9.680e-02   2.694 0.007921 **
## Internet     7.167e-02  2.635e-02   2.720 0.007340 **
## BirthRate   -7.248e-01  6.791e-02 -10.673  < 2e-16 ***
## ElderlyPop  -4.429e-01  1.398e-01  -3.169 0.001878 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.734 on 141 degrees of freedom
## Multiple R-squared:  0.7914, Adjusted R-squared:  0.7825
## F-statistic: 89.16 on 6 and 141 DF,  p-value: < 2.2e-16
```

```
anova(model5)
```

```
## Analysis of Variance Table
##
## Response: LifeExpectancy
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## LandArea      1   20.0    20.0   0.8904  0.346991
## Rural         1 7009.6  7009.6 312.7269 < 2.2e-16 ***
## Health        1  648.1   648.1  28.9134 3.064e-07 ***
## Internet      1 1594.8  1594.8  71.1514 3.543e-14 ***
## BirthRate     1 2493.4  2493.4 111.2401 < 2.2e-16 ***
## ElderlyPop    1  225.1   225.1  10.0405  0.001878 **
## Residuals   141 3160.4    22.4
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# transform the response variable using the lambda from the boxcox plot
life_ex <- countries_df$LifeExpectancy
trans_longevity <- ((life_ex ^ 1.8) - 1)/1.8
final_model <- lm(trans_longevity~LandArea+Rural+Health+Internet+BirthRate+ElderlyPop+Cell,
                  data = countries_df)
# summarize the final transformed model
summary(final_model)
```

```
##
## Call:
## lm(formula = trans_longevity ~ LandArea + Rural + Health + Internet +
##     BirthRate + ElderlyPop + Cell, data = countries_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -425.65  -74.85    9.60   80.69  289.84
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.547e+03  9.064e+01  17.064  < 2e-16 ***
## LandArea    -8.007e-06  5.492e-06  -1.458  0.14710
## Rural       -2.164e+00  6.882e-01  -3.144  0.00203 **
## Health       7.650e+00  2.618e+00   2.922  0.00405 **
## Internet     2.310e+00  7.163e-01   3.225  0.00157 **
## BirthRate   -1.816e+01  1.975e+00  -9.192 4.79e-16 ***
## ElderlyPop  -1.048e+01  3.825e+00  -2.739  0.00697 **
## Cell         3.999e-01  3.438e-01   1.163  0.24680
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 128 on 140 degrees of freedom
## Multiple R-squared:  0.8088, Adjusted R-squared:  0.7992
## F-statistic:  84.6 on 7 and 140 DF,  p-value: < 2.2e-16
```