Regression Model: Average Global Life Expectancy Prediction

Gurdeep Panag (30101520), Harjot Dhaliwal (30051859), Lukas Escoda (30211208), Shabbir Khandwala (30219011)

University of Calgary

DATA 603: Statistical Modelling with Data

Instructor: Dr. Long

December 7th, 2023

Introduction

Our project centers around utilizing data on average life expectancy from the World Health Organization, spanning 179 countries between 2000 and 2015. Our objective is to construct a predictive regression model, considering various factors, to offer insights into enhancing a country's average life expectancy. The ultimate aim is to empower nations with proactive measures, enabling them to address critical factors and potentially avert national crises, ultimately fostering longer and healthier lives.

Our progress intends to identify pivotal variables impacting average life expectancy, crafting models that enable leaders to make informed, data-driven decisions. Our goal is to develop a regression model with high prediction accuracy, emphasizing relevant variables and extending predictions to unseen data. Additionally, we aim to uncover insights on the variables most influencing a nation's life expectancy, proposing strategies for optimization.

In terms of statistical methodology, we plan to leverage various visual plots (Ex. Residual Plots) to validate model assumptions and depict life expectancy trends globally. Through regression techniques, we aim to select an accurate prediction model, verify assumptions, and provide valuable insights for world leaders.

This project holds personal significance for us, particularly those with roots in developing nations. We aspire to contribute to the improvement of living conditions in these regions, promoting longer and healthier lives to stimulate overall national growth. We firmly believe that enhancing living conditions globally will yield mutual benefits for people worldwide, fostering a stronger and more interconnected world.

Dataset

The dataset, sourced from Kaggle (Lasha., 2023), comprises information from 179 countries spanning the years 2000-2015 and encompasses 21 features. Key variables such as Population, GDP, and Life Expectancy were inputted using World Bank data. Details on vaccinations, alcohol consumption, BMI, HIV incidents, mortality rates, and thinness were derived from World Health Organization datasets, while schooling

information was obtained from Our World in Data, a project of the University of Oxford. The dataset introduces a variable classifying countries as Developed or Developing, adhering to definitions provided by the World Trade Organization and the United Nations.

Key features include:

1. **Country:** List of the 179 countries.
2. **Region:** The distribution of 179 countries across 9 regions, including Africa, Asia, Oceania, European Union, Rest of Europe, etc.
3. **Year:** Observed years ranging from 2000 to 2015.
4. **Infant Deaths:** Represents infant deaths per 1000 population.
5. **Under-Five Years Old Deaths:** Represents deaths of children under five years old per 1000 population.
6. **Adult Mortality:** Represents deaths of adults per 1000 population.
7. **Alcohol Consumption:** Represents recorded alcohol consumption in liters of pure alcohol per capita for individuals aged 15 and above.
8. **Hepatitis B:** Represents the percentage of coverage of Hepatitis B (HepB3) immunization among 1-year-olds.
9. **Measles:** Represents the percentage of coverage of Measles-containing vaccine first dose (MCV1) immunization among 1-year-olds.
10. **BMI:** A measure of nutritional status in adults, calculated as a person's weight in kilograms divided by the square of their height in meters (kg/m2).
11. **Polio:** Represents the percentage of coverage of Polio (Pol3) immunization among 1-year-olds.
12. **Diphtheria:** Represents the percentage of coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1-year-olds.
13. **Incidents_HIV:** Incidents of HIV per 1000 population aged 15-49.
14. **GDP per Capita:** GDP per capita in current USD.
15. **Population (Millions):** Total population in millions.
16. **Thinness (10-19 Years):** Prevalence of thinness among adolescents aged 10-19 years (BMI < -2 standard deviations below the median).
17. **Thinness (5-9 Years):** Prevalence of thinness among children aged 5-9 years (BMI < -2 standard deviations below the median).
18. **Schooling:** Average years that people aged 25 and above spent in formal education.
19. **Economy Status (Developed):** Developed country.
20. **Economy Status (Developing):** Developing country.
21. **Life Expectancy:** Average life expectancy of both genders in different years from 2010 to 2015.

Methodology

We aim to streamline our model selection process by employing a systematic approach. Initially, we will utilize forward selection, followed by backward selection and stepwise methods to identify the most impactful variables. Subsequently, we will construct a new model incorporating consistently selected variables and conduct a best subset model selection. This strategy optimizes computational resources by eliminating insignificant variables first. To further refine our model, we will explore higher order and interaction terms. This

approach ensures that we consider the best variables identified through various methods and assess if there are superior model subsets. Our goal is to strike a balance between computational efficiency and a thorough search for the best model, ensuring a robust selection process.

The project unfolds in two main phases. First, we conduct an extensive model selection process, followed by the identification of interaction terms and higher order terms, and then confirming adherence to regression assumptions. The challenge lies in the subjective nature of model selection, requiring us to leverage collective knowledge. There is no definitive “wrong” answer in this step, emphasizing the need for a careful and informed decision-making process. In the event that assumptions are not met, we would utlize corrective actions such as variable deletion for multicollinearity, log transformation, or box-cox transformation.

Our work is divided into two parts: Lukas and Shabbir will focus on selecting the best model, while Gurdeep and Harjot will enhance the model with higher order terms, interaction terms, and assess regression assumptions. Concluded with a joint effort to finalize the write up and report.

Results

$$\frac{life\_expectancy^{3.33}-1}{3.33} = -774459.69 + 122320.49xschooling$$
$$+ 2257.98x_{polio} + 10.96x_{gdp/capita} + 36066.39x_{bmi} - 150716.01x_{incidentshiv} -$$
$$0.00016x^2_{gdp/capita} + 0.0000000008x^3_{gdp/capita} + 4852.97x_{incidentshiv} * x_{bmi} -$$
$$4396.57x_{schooling} * x_{bmi}$$

Our exhaustive regression model selection process unveiled crucial predictors significantly linked to life expectancy. Notably, variables such as schooling, Polio immunization coverage, GDP per Capita, BMI, and incidents of HIV emerged as influential factors. The model revealed that the average number of years of schooling is associated with increase in life expectancy, advocating for investments in education to foster longer lives. Similarly, higher Polio immunization coverage exhibited a positive correlation with life expectancy, suggesting fortified immunization programs for improved life spans. Surprisingly, the relationship between GDP per Capita and life expectancy showcased a cubic trend. Encouraging healthy BMI ranges and effective management of HIV incidents were also pivotal, showcasing positive correlations with longer life expectancy. This nuanced understanding challenges the traditional belief in continual economic growth as the sole driver of increased life spans. Our model’s insights offer strategic guidance to nations, advocating for a holistic approach encompassing education, healthcare initiatives, balanced economic growth, and disease management to foster longer and healthier lives among populations.

Conclusion

The approach undertaken for our project exhibits promise in uncovering influential factors impacting life expectancy across nations. By employing a systematic model selection process, we identified key predictors, including education, healthcare initiatives, economic indicators, and disease management, that significantly influence life expectancy. However, considering the complexity of socio-economic factors affecting longevity, augmenting our approach with a more expansive dataset, encompassing a wider array of socio-cultural

aspects, could offer a more comprehensive understanding. Additionally, integrating advanced machine learning techniques, such as ensemble methods or deep learning architectures, might provide more nuanced insights into the intricate interplay of factors affecting life expectancy. While our current approach is robust, further exploration involving broader datasets and advanced methodologies could enhance the depth and accuracy of our predictions.

In considering future endeavors, several avenues present themselves for further exploration and refinement. First and foremost, expanding our dataset to incorporate cultural, environmental, and policy-related variables could enrich our understanding of life expectancy determinants. Exploring regional or local-level data to capture nuances within countries could provide more targeted strategies for enhancing life expectancy at a more granular level. Moreover, validating our findings through longitudinal studies or by incorporating real-time data streams could offer dynamic insights into evolving factors influencing life expectancy. Additionally, collaborating with public health organizations or governmental bodies to implement and assess the impact of interventions based on our findings could provide tangible validations and foster the practical application of our research. Continual refinement of our predictive models, validation of assumptions, and addressing potential confounding variables are critical for establishing robust and reliable recommendations for policymakers and global health practitioners.

References

Lasha., 2023. Life expectancy (WHO) fixed. Kaggle. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated (https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated)

Appendix

Model Selection

Let's begin by importing the dataset, focusing exclusively on the year 2015 (in order to preserve independence of the error terms). Furthermore, we would like to use the rest of the data to test against our final regression model.

```
life.df = read.csv("Life-Expectancy-Data-Updated.csv")
life = filter(life.df, Year == 2015)
head(life, 4)
```

	Country <chr>	Region <chr>	Y… <int>	Infant_deaths <dbl>	Under_five_deaths <dbl>	Adult_r
1	Turkiye	Middle East	2015	11.1	13.0	1
2	Spain	European Union	2015	2.7	3.3	
3	Russian Federation	Rest of Europe	2015	6.6	8.2	2

4	Cameroon	Africa	2015	57.0	88.0	3

4 rows | 1-7 of 22 columns

Now we can establish a maximum first order model. We intentionally excluded country, region, and year from our maximum first order model as we wanted to preserve independence of the error terms. Furthermore, we do not want factors that are outside of the control of stakeholders, to be included in our final model. In addition we removed all mortality-related variables as these variables are highly correlated with life-expectancy, therefore, overshadowing other variables, ultimately leading to their insignificance in a hypothetical final model. This kind of model would not allow us to derive any valuable insights for our original guiding question to provide people with as all of these are “death” related factors. Therefore, this is the appropriate maximum first order model we would start to work with.

```
basemodel = lm(Life_expectancy ~ Alcohol_consumption + Hepatitis_B + Measles + BMI +
Polio + Diphtheria + Incidents_HIV + GDP_per_capita + Population_mln + Thinness_ten_n
ineteen_years + Thinness_five_nine_years + Schooling + factor(Economy_status_Develope
d), data = life)

summary(basemodel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption + Hepatitis_B +
##     Measles + BMI + Polio + Diphtheria + Incidents_HIV + GDP_per_capita +
##     Population_mln + Thinness_ten_nineteen_years + Thinness_five_nine_years +
##     Schooling + factor(Economy_status_Developed), data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2751  -2.2856   0.2349   2.8534   7.9669
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.414e+01  4.896e+00   6.973 7.10e-11 ***
## Alcohol_consumption              -3.792e-02  1.164e-01  -0.326 0.744949
## Hepatitis_B                      -2.951e-02  6.561e-02  -0.450 0.653499
## Measles                           4.306e-02  2.257e-02   1.908 0.058082 .
## BMI                               5.922e-01  1.886e-01   3.140 0.002005 **
## Polio                             1.358e-01  6.099e-02   2.226 0.027363 *
## Diphtheria                        3.274e-02  7.617e-02   0.430 0.667893
## Incidents_HIV                    -1.339e+00  1.783e-01  -7.506 3.61e-12 ***
## GDP_per_capita                    9.149e-05  2.278e-05   4.016 8.97e-05 ***
## Population_mln                    2.752e-03  1.989e-03   1.384 0.168308
## Thinness_ten_nineteen_years      -5.561e-01  3.027e-01  -1.838 0.067931 .
## Thinness_five_nine_years          5.492e-01  3.031e-01   1.812 0.071817 .
## Schooling                         6.880e-01  1.776e-01   3.873 0.000155 ***
## factor(Economy_status_Developed)1 1.952e+00  1.183e+00   1.651 0.100649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.638 on 165 degrees of freedom
## Multiple R-squared:  0.8001, Adjusted R-squared:  0.7843
## F-statistic: 50.79 on 13 and 165 DF,  p-value: < 2.2e-16
```

Now, we can perform a step-wise procedure, backward elimination procedure and forward selection procedure in order to see which variables to keep and remove.

```
print("STEPWISE")
```

```
## [1] "STEPWISE"
```

```
summary(ols_step_both_p(basemodel, pent = 0.05, prem = 0.1, details = FALSE)$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7482  -2.5039   0.2098   2.5154   7.1598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.803e+01  4.026e+00   9.445  < 2e-16 ***
## Schooling       8.967e-01  1.425e-01   6.294 2.47e-09 ***
## Incidents_HIV  -1.440e+00  1.731e-01  -8.319 2.55e-14 ***
## Polio           1.528e-01  2.447e-02   6.243 3.21e-09 ***
## GDP_per_capita  1.101e-04  1.961e-05   5.612 7.80e-08 ***
## BMI             4.667e-01  1.593e-01   2.930  0.00385 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 173 degrees of freedom
## Multiple R-squared:  0.7866, Adjusted R-squared:  0.7804
## F-statistic: 127.5 on 5 and 173 DF,  p-value: < 2.2e-16
```

```
print("FORWARD")
```

```
## [1] "FORWARD"
```

```
summary(ols_step_forward_p(basemodel, pent = 0.05, details= FALSE)$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7482  -2.5039   0.2098   2.5154   7.1598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.803e+01  4.026e+00   9.445  < 2e-16 ***
## Schooling       8.967e-01  1.425e-01   6.294 2.47e-09 ***
## Incidents_HIV  -1.440e+00  1.731e-01  -8.319 2.55e-14 ***
## Polio           1.528e-01  2.447e-02   6.243 3.21e-09 ***
## GDP_per_capita  1.101e-04  1.961e-05   5.612 7.80e-08 ***
## BMI             4.667e-01  1.593e-01   2.930  0.00385 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 173 degrees of freedom
## Multiple R-squared:  0.7866, Adjusted R-squared:  0.7804
## F-statistic: 127.5 on 5 and 173 DF,  p-value: < 2.2e-16
```

```
print("BACKWARD")
```

```
## [1] "BACKWARD"
```

```
summary(ols_step_backward_p(basemodel, prem = 0.05, details = FALSE)$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7482  -2.5039   0.2098   2.5154   7.1598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.803e+01  4.026e+00   9.445  < 2e-16 ***
## BMI             4.667e-01  1.593e-01   2.930  0.00385 **
## Polio           1.528e-01  2.447e-02   6.243 3.21e-09 ***
## Incidents_HIV  -1.440e+00  1.731e-01  -8.319 2.55e-14 ***
## GDP_per_capita  1.101e-04  1.961e-05   5.612 7.80e-08 ***
## Schooling       8.967e-01  1.425e-01   6.294 2.47e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 173 degrees of freedom
## Multiple R-squared:  0.7866, Adjusted R-squared:  0.7804
## F-statistic: 127.5 on 5 and 173 DF,  p-value: < 2.2e-16
```

All of the selection procedures tell us to keep the same variables; schooling, incidents_hiv, polio, gdp_per_capita and bmi.

Therefore, comparing our best reduced model against the base model using a partial F-test we see that:

```
reducedmodel = ols_step_both_p(basemodel, pent = 0.05, prem = 0.1, details = FALSE)$m
odel

anova(basemodel, reducedmodel)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	165	2183.203	*NA*	*NA*	*NA*	*NA*
2	173	2330.671	-8	-147.4685	1.393155	0.2028915

2 rows

H_0 : No significant variables excluded from the mode. H_A : At least one significant variable excluded from the model.

From the partial F-test, we can see that the P value is greater than 0.05, meaning that the reduced model is not missing any significant variables when compared to the base model.

Now we can perform all possible regression model selection on these 5 variables only, in order to preserve compute resources, and see if any smaller models are still viable at this stage.

```
reducedmodel = lm(Life_expectancy ~ BMI + Polio + Incidents_HIV + GDP_per_capita + Sc
hooling, data = life)

all_possible_regression = ols_step_best_subset(reducedmodel, details = FALSE)
all_possible_regression
```

min...	n	predictors	rsquare	adj
<int>	<int>	<chr>	<dbl>	<dbl
5	1	1 Schooling	0.5895824	0.587263
14	2	2 Incidents_HIV Schooling	0.6931283	0.689641
23	3	3 Polio Incidents_HIV Schooling	0.7416772	0.737248
30	4	4 Polio Incidents_HIV GDP_per_capita Schooling	0.7759667	0.770816
31	5	5 BMI Polio Incidents_HIV GDP_per_capita Schooling	0.7865552	0.780386

5 rows | 1-7 of 15 columns

We can see here that according to the smaller values for AIC and Cp as well as the larger adjusted r squared value, the best model to choose would be the one with all 5 of these variables, as these differences in the selection criteria are significant enough to justify the increase in model complexity.

Interaction Terms

Now, we can go ahead and check for interaction terms:

```
intmodel = lm(Life_expectancy ~ (BMI + Polio + Incidents_HIV + GDP_per_capita + Schoo
ling) ^ 2, data = life)

summary(intmodel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ (BMI + Polio + Incidents_HIV +
##     GDP_per_capita + Schooling)^2, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4492 -1.7766  0.1193  2.0747  7.7754
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3.739e+01  2.451e+01  -1.525 0.129126
## BMI                            3.641e+00  1.081e+00   3.369 0.000941 ***
## Polio                          7.251e-01  2.904e-01   2.497 0.013532 *
## Incidents_HIV                 -1.174e+01  3.264e+00  -3.597 0.000427 ***
## GDP_per_capita                 1.846e-03  6.175e-04   2.990 0.003221 **
## Schooling                      4.434e+00  1.677e+00   2.644 0.009001 **
## BMI:Polio                     -2.468e-02  1.273e-02  -1.939 0.054259 .
## BMI:Incidents_HIV              3.304e-01  1.121e-01   2.947 0.003683 **
## BMI:GDP_per_capita            -5.623e-06  1.184e-05  -0.475 0.635408
## BMI:Schooling                 -1.694e-01  5.718e-02  -2.962 0.003512 **
## Polio:Incidents_HIV            2.866e-02  2.188e-02   1.310 0.192031
## Polio:GDP_per_capita          -1.229e-05  4.636e-06  -2.651 0.008827 **
## Polio:Schooling                8.369e-03  9.918e-03   0.844 0.400004
## Incidents_HIV:GDP_per_capita   3.397e-05  7.273e-05   0.467 0.641035
## Incidents_HIV:Schooling       -1.255e-01  1.453e-01  -0.864 0.388822
## GDP_per_capita:Schooling      -3.303e-05  1.199e-05  -2.755 0.006532 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.251 on 163 degrees of freedom
## Multiple R-squared:  0.8422, Adjusted R-squared:  0.8277
## F-statistic:    58 on 15 and 163 DF,  p-value: < 2.2e-16
```

It appears that there are only 4 significant interaction terms here, as the corresponding p-values are below our threshold of 0.05, therefore we can remove the insignificant ones.

```
reducedintmodel = lm(Life_expectancy ~ BMI + Polio + Incidents_HIV + GDP_per_capita +
Schooling + BMI:Incidents_HIV + BMI:Schooling + Polio:GDP_per_capita + GDP_per_capit
a:Schooling, data = life)

summary(reducedintmodel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ BMI + Polio + Incidents_HIV +
##     GDP_per_capita + Schooling + BMI:Incidents_HIV + BMI:Schooling +
##     Polio:GDP_per_capita + GDP_per_capita:Schooling, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4592 -1.9704  0.1613  2.4215  7.2540
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 8.541e-01  9.973e+00   0.086 0.931854
## BMI                         1.887e+00  4.091e-01   4.613 7.82e-06 ***
## Polio                       1.811e-01  2.662e-02   6.800 1.72e-10 ***
## Incidents_HIV              -9.914e+00  2.667e+00  -3.717 0.000274 ***
## GDP_per_capita              1.424e-03  3.781e-04   3.766 0.000229 ***
## Schooling                   6.435e+00  1.175e+00   5.476 1.55e-07 ***
## BMI:Incidents_HIV           3.246e-01  1.037e-01   3.129 0.002067 **
## BMI:Schooling              -2.222e-01  4.754e-02  -4.673 6.03e-06 ***
## Polio:GDP_per_capita       -1.077e-05  3.617e-06  -2.978 0.003324 **
## GDP_per_capita:Schooling -2.216e-05  1.038e-05  -2.136 0.034132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.266 on 169 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.8262
## F-statistic:    95 on 9 and 169 DF,  p-value: < 2.2e-16
```

This is now our best reduced interaction model, now we can perform the partial F-test based on the full interaction model vs. the reduced interaction model in order to confirm if we only removed insignificant terms:

```
anova(intmodel, reducedintmodel)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	163	1723.100	*NA*	*NA*	*NA*	*NA*
2	169	1802.192	-6	-79.09167	1.246971	0.2850731

2 rows

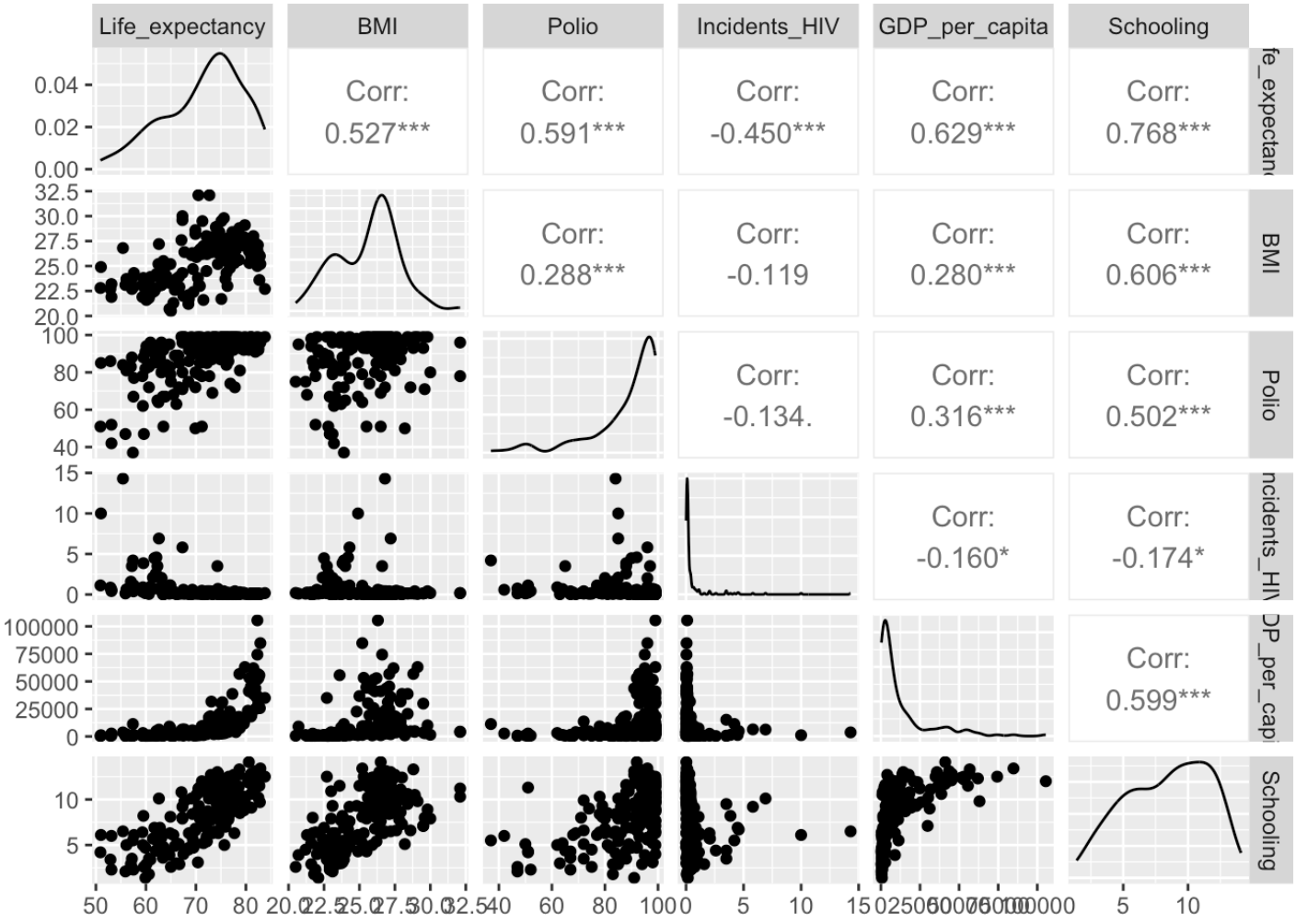
$H_0: $ No significant variables excluded from the model. $H_A: $ At least one significant variable excluded from the model.

From the partial F-test, we can see that there is no significant interaction terms left out of our new model, since the p-value is greater than 0.05.

Higher Order Terms

Now, we can check for any higher-order terms:

```
ggpairs(life[, c("Life_expectancy", "BMI", "Polio", "Incidents_HIV", "GDP_per_capit
a", "Schooling")])
```



From these graphs, we can see that there may be a non-linear relationship with gdp_per_capita due to the curved nature of the scatter plot. We can add a higher order term for this variable to our model:

```
reducedpwrintmodel = lm(Life_expectancy ~ Schooling + Polio + GDP_per_capita + BMI +
Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) + BMI*Incidents_HIV + BMI*School
ing + Polio*GDP_per_capita + GDP_per_capita*Schooling, data = life)

summary(reducedpwrintmodel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Schooling + Polio + GDP_per_capita +
##     BMI + Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) +
##     BMI * Incidents_HIV + BMI * Schooling + Polio * GDP_per_capita +
##     GDP_per_capita * Schooling, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4394 -1.8324  0.1727  2.3561  7.2847
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.172e+00  9.817e+00  -0.119 0.905138
## Schooling                  6.663e+00  1.156e+00   5.762 3.88e-08 ***
## Polio                      1.663e-01  2.668e-02   6.234 3.55e-09 ***
## GDP_per_capita             1.018e-03  3.999e-04   2.545 0.011819 *
## BMI                        2.050e+00  4.060e-01   5.050 1.14e-06 ***
## Incidents_HIV             -9.435e+00  2.624e+00  -3.596 0.000425 ***
## I(GDP_per_capita^2)       -1.686e-09  6.186e-10  -2.726 0.007097 **
## BMI:Incidents_HIV          3.072e-01  1.020e-01   3.011 0.003004 **
## Schooling:BMI             -2.373e-01  4.699e-02  -5.051 1.14e-06 ***
## Polio:GDP_per_capita      -6.783e-06  3.841e-06  -1.766 0.079202 .
## Schooling:GDP_per_capita -9.813e-06  1.115e-05  -0.880 0.379954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.205 on 168 degrees of freedom
## Multiple R-squared:  0.8419, Adjusted R-squared:  0.8325
## F-statistic: 89.49 on 10 and 168 DF,  p-value: < 2.2e-16
```

Now, we can check the cubic term.

```
reducedpwrintmodel = lm(Life_expectancy ~ Schooling + Polio + GDP_per_capita + BMI +
Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) + I(GDP_per_capita^3) + BMI*Inci
dents_HIV + BMI*Schooling + Polio*GDP_per_capita + GDP_per_capita*Schooling, data = l
ife)

summary(reducedpwrintmodel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Schooling + Polio + GDP_per_capita +
##     BMI + Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) +
##     I(GDP_per_capita^3) + BMI * Incidents_HIV + BMI * Schooling +
##     Polio * GDP_per_capita + GDP_per_capita * Schooling, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3924 -1.9295  0.2178  2.0508  7.2502
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.726e+00  9.681e+00  -0.178 0.858709
## Schooling                 6.614e+00  1.140e+00   5.801 3.22e-08 ***
## Polio                     1.651e-01  2.631e-02   6.276 2.89e-09 ***
## GDP_per_capita            1.088e-03  3.953e-04   2.751 0.006594 **
## BMI                       2.097e+00  4.007e-01   5.234 4.93e-07 ***
## Incidents_HIV            -9.010e+00  2.593e+00  -3.475 0.000651 ***
## I(GDP_per_capita^2)      -8.445e-09  2.861e-09  -2.951 0.003618 **
## I(GDP_per_capita^3)       4.717e-14  1.951e-14   2.418 0.016693 *
## BMI:Incidents_HIV         2.910e-01  1.008e-01   2.887 0.004408 **
## Schooling:BMI            -2.419e-01  4.637e-02  -5.216 5.35e-07 ***
## Polio:GDP_per_capita     -6.829e-06  3.786e-06  -1.804 0.073100 .
## Schooling:GDP_per_capita  4.173e-06  1.242e-05   0.336 0.737257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.16 on 167 degrees of freedom
## Multiple R-squared:  0.8473, Adjusted R-squared:  0.8372
## F-statistic: 84.23 on 11 and 167 DF,  p-value: < 2.2e-16
```

Now we can check the fourth term:

```
reducedpwrintmodel = lm(Life_expectancy ~ Schooling + Polio + GDP_per_capita + BMI +
Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) + I(GDP_per_capita^3) + I(GDP_pe
r_capita^4) + BMI*Incidents_HIV + BMI*Schooling + Polio*GDP_per_capita + GDP_per_capi
ta*Schooling, data = life)

summary(reducedpwrintmodel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Schooling + Polio + GDP_per_capita +
##     BMI + Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) +
##     I(GDP_per_capita^3) + I(GDP_per_capita^4) + BMI * Incidents_HIV +
##     BMI * Schooling + Polio * GDP_per_capita + GDP_per_capita *
##     Schooling, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3723 -1.9522  0.3199  2.0892  7.1778
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.662e-01  9.777e+00  -0.037 0.970172
## Schooling                 6.499e+00  1.146e+00   5.671 6.16e-08 ***
## Polio                     1.596e-01  2.690e-02   5.932 1.69e-08 ***
## GDP_per_capita            1.088e-03  3.953e-04   2.752 0.006589 **
## BMI                       2.057e+00  4.027e-01   5.109 8.82e-07 ***
## Incidents_HIV            -8.767e+00  2.604e+00  -3.366 0.000947 ***
## I(GDP_per_capita^2)      -1.579e-08  7.923e-09  -1.994 0.047833 *
## I(GDP_per_capita^3)       1.765e-13  1.315e-13   1.343 0.181244
## I(GDP_per_capita^4)      -6.882e-19  6.918e-19  -0.995 0.321262
## BMI:Incidents_HIV         2.817e-01  1.012e-01   2.783 0.006010 **
## Schooling:BMI            -2.386e-01  4.649e-02  -5.131 7.96e-07 ***
## Polio:GDP_per_capita     -5.665e-06  3.963e-06  -1.429 0.154824
## Schooling:GDP_per_capita  6.184e-06  1.258e-05   0.491 0.623750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.16 on 166 degrees of freedom
## Multiple R-squared:  0.8482, Adjusted R-squared:  0.8372
## F-statistic: 77.29 on 12 and 166 DF,  p-value: < 2.2e-16
```

We can see that the fourth power term is insignificant, therefore we stick with the cubic model.

```
reducedpwrintmodel = lm(Life_expectancy ~ Schooling + Polio + GDP_per_capita + BMI +
Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) + I(GDP_per_capita^3) + BMI*Inci
dents_HIV + BMI*Schooling + Polio*GDP_per_capita + GDP_per_capita*Schooling, data = l
ife)

summary(reducedpwrintmodel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Schooling + Polio + GDP_per_capita +
##     BMI + Incidents_HIV + GDP_per_capita + I(GDP_per_capita^2) +
##     I(GDP_per_capita^3) + BMI * Incidents_HIV + BMI * Schooling +
##     Polio * GDP_per_capita + GDP_per_capita * Schooling, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3924 -1.9295  0.2178  2.0508  7.2502
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.726e+00  9.681e+00  -0.178 0.858709
## Schooling                 6.614e+00  1.140e+00   5.801 3.22e-08 ***
## Polio                     1.651e-01  2.631e-02   6.276 2.89e-09 ***
## GDP_per_capita            1.088e-03  3.953e-04   2.751 0.006594 **
## BMI                       2.097e+00  4.007e-01   5.234 4.93e-07 ***
## Incidents_HIV            -9.010e+00  2.593e+00  -3.475 0.000651 ***
## I(GDP_per_capita^2)      -8.445e-09  2.861e-09  -2.951 0.003618 **
## I(GDP_per_capita^3)       4.717e-14  1.951e-14   2.418 0.016693 *
## BMI:Incidents_HIV         2.910e-01  1.008e-01   2.887 0.004408 **
## Schooling:BMI            -2.419e-01  4.637e-02  -5.216 5.35e-07 ***
## Polio:GDP_per_capita     -6.829e-06  3.786e-06  -1.804 0.073100 .
## Schooling:GDP_per_capita  4.173e-06  1.242e-05   0.336 0.737257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.16 on 167 degrees of freedom
## Multiple R-squared:  0.8473, Adjusted R-squared:  0.8372
## F-statistic: 84.23 on 11 and 167 DF,  p-value: < 2.2e-16
```

We can see now that the interaction terms Polio:GDP_per_capita and Schooling:GDP_per_capita are insignificant, therefore we can remove them.

```
reducedpwrintmodel = lm(Life_expectancy ~ Schooling + Polio + GDP_per_capita + I(GDP_
per_capita^2) + I(GDP_per_capita^3) + BMI + Incidents_HIV + BMI*Incidents_HIV + BMI*S
chooling, data = life)

summary(reducedpwrintmodel)
```

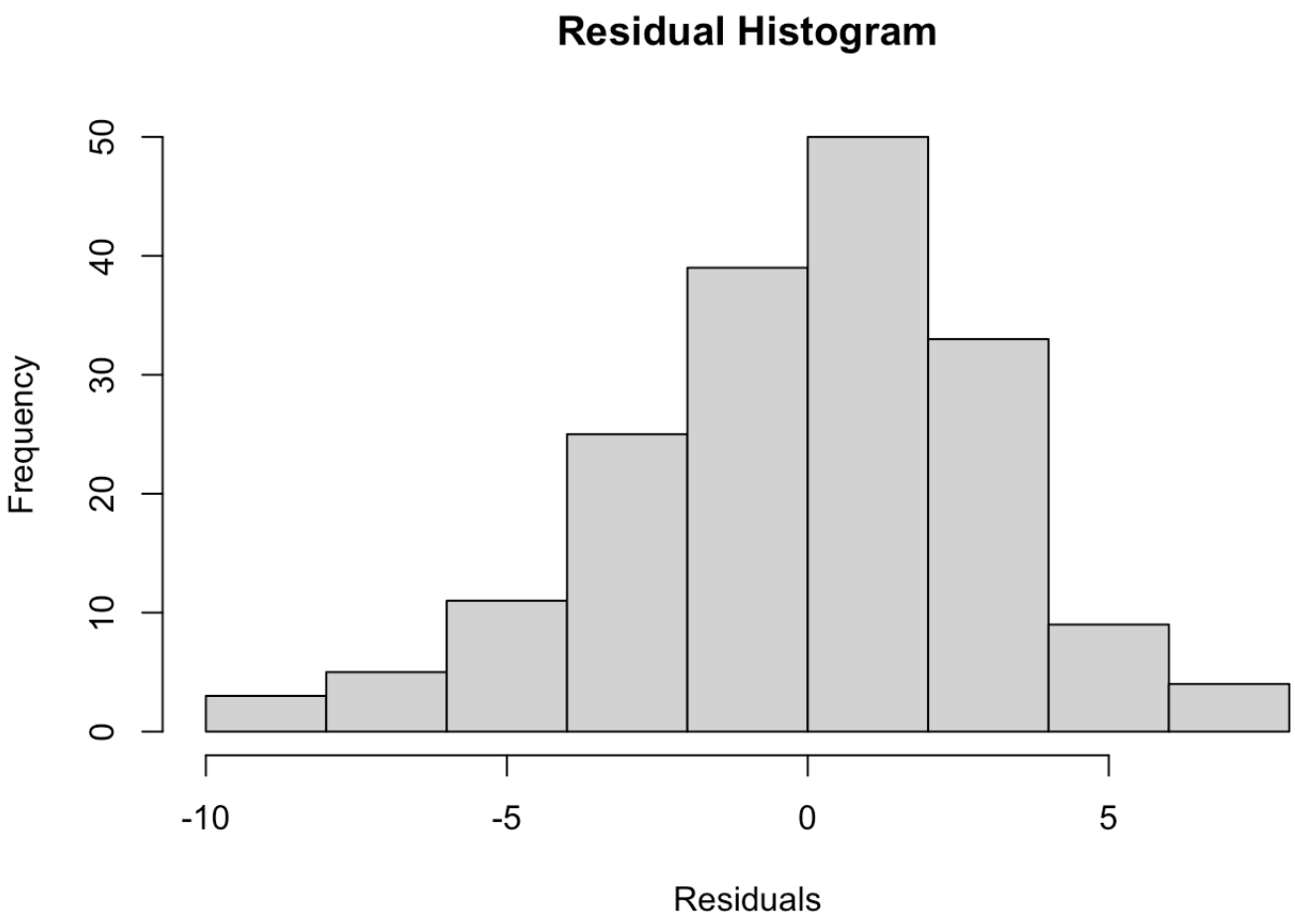
```
##
## Call:
## lm(formula = Life_expectancy ~ Schooling + Polio + GDP_per_capita +
##     I(GDP_per_capita^2) + I(GDP_per_capita^3) + BMI + Incidents_HIV +
##     BMI * Incidents_HIV + BMI * Schooling, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3365 -1.9762  0.1285  2.1659  7.0731
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.999e+00  9.095e+00   0.220  0.82627
## Schooling             6.506e+00  1.122e+00   5.800 3.18e-08 ***
## Polio                 1.356e-01  2.138e-02   6.341 2.00e-09 ***
## GDP_per_capita        4.699e-04  8.930e-05   5.262 4.27e-07 ***
## I(GDP_per_capita^2) -7.821e-09  2.435e-09  -3.212  0.00158 **
## I(GDP_per_capita^3)  4.074e-14  1.726e-14   2.360  0.01940 *
## BMI                   2.029e+00  3.730e-01   5.440 1.85e-07 ***
## Incidents_HIV        -8.627e+00  2.591e+00  -3.330  0.00107 **
## BMI:Incidents_HIV     2.773e-01  1.008e-01   2.752  0.00658 **
## Schooling:BMI        -2.343e-01  4.468e-02  -5.243 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.176 on 169 degrees of freedom
## Multiple R-squared:  0.8439, Adjusted R-squared:  0.8356
## F-statistic: 101.5 on 9 and 169 DF,  p-value: < 2.2e-16
```

We can conclude now, that this is our best final model that we can test the assumptions for.

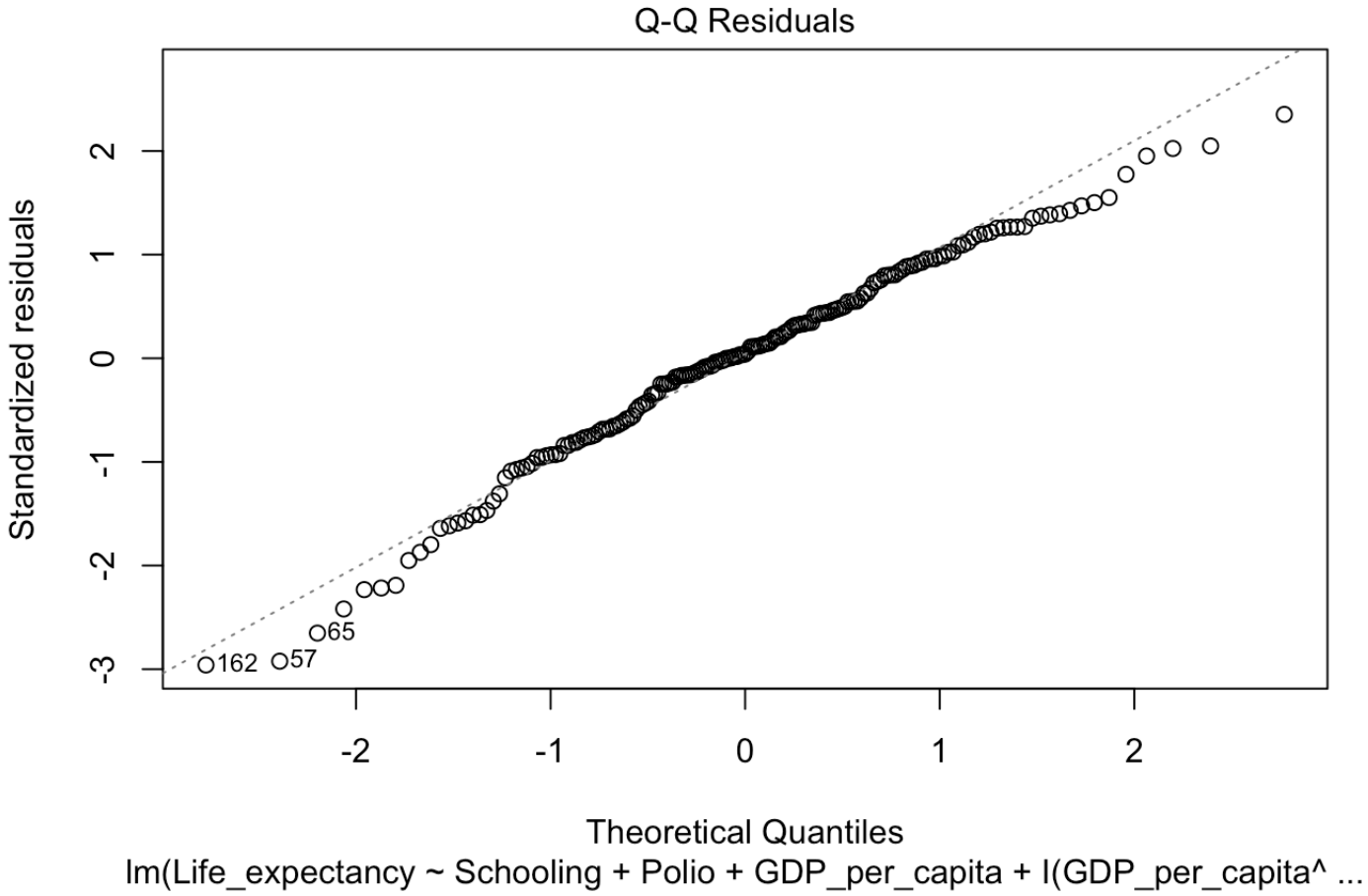
Assumption Testing

Normality of the Residuals:

```
hist(residuals(reducedpwrintmodel), main = "Residual Histogram", xlab = "Residuals")
```




```
plot(reducedpwrintmodel, which = 2)
```




```
shapiro.test(residuals(reducedpwrintmodel))
```

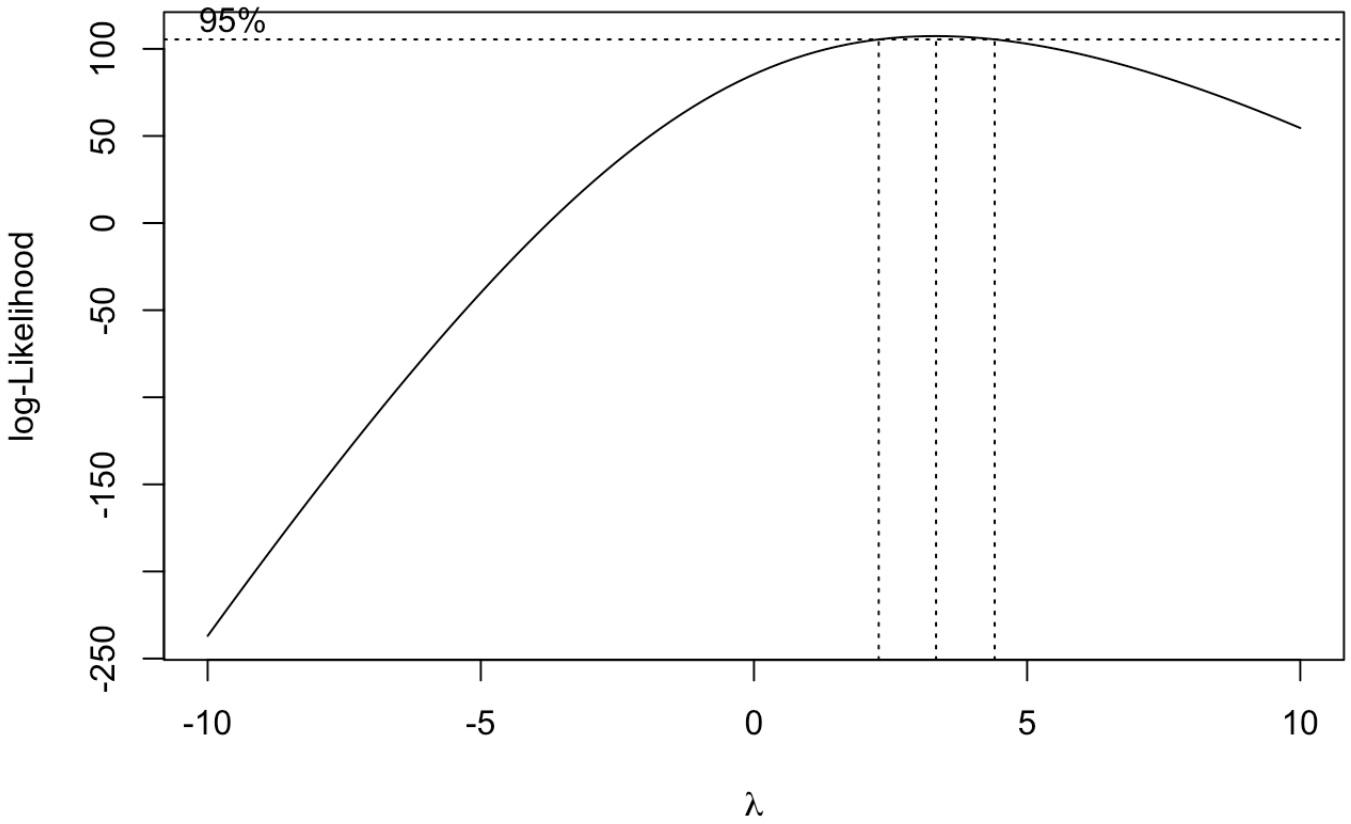
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(reducedpwrintmodel)
## W = 0.98207, p-value = 0.02112
```

$H_0: $ Residuals are normally distributed. $H_A: $ Residuals are not normally distributed.

As we can see, the Shapiro-Wilk fails as there is a p-value less than 0.05, therefore rejecting the null hypothesis and saying that our condition of normality of the residuals does not hold. In this case, we can go ahead and try to do a box-cox transformation to check if this helps meet this assumption:

Box-Cox Transformation:

```
bc = boxcox(reducedpwrintmodel, lambda = seq(-10, 10), data = life)
```




```
best.lambda = bc$x[which(bc$y == max(bc$y))]
best.lambda
```

```
## [1] 3.333333
```

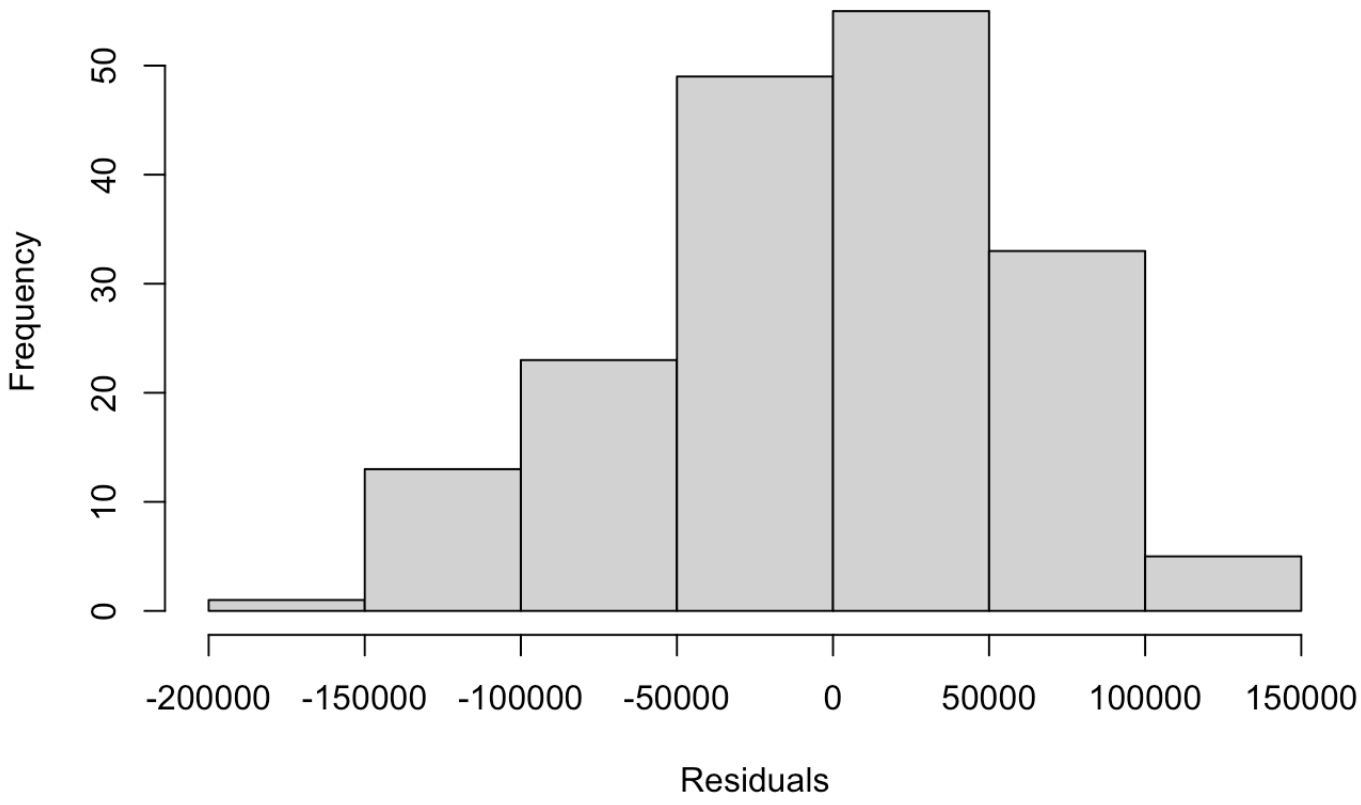
```
boxcoxmodel = lm((Life_expectancy^best.lambda - 1)/best.lambda ~ Schooling + Polio +
GDP_per_capita + I(GDP_per_capita^2) + I(GDP_per_capita^3) + BMI + Incidents_HIV + BM
I*Incidents_HIV + BMI*Schooling, data = life)

summary(boxcoxmodel)
```

```
##
## Call:
## lm(formula = (Life_expectancy^best.lambda - 1)/best.lambda ~
##     Schooling + Polio + GDP_per_capita + I(GDP_per_capita^2) +
##         I(GDP_per_capita^3) + BMI + Incidents_HIV + BMI * Incidents_HIV +
##         BMI * Schooling, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -153132  -38254    3496   41796  143326
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -7.745e+05  1.793e+05  -4.318 2.67e-05 ***
## Schooling            1.223e+05  2.212e+04   5.530 1.19e-07 ***
## Polio                2.258e+03  4.216e+02   5.356 2.74e-07 ***
## GDP_per_capita       1.096e+01  1.761e+00   6.227 3.65e-09 ***
## I(GDP_per_capita^2) -1.673e-04  4.801e-05  -3.484 0.000629 ***
## I(GDP_per_capita^3)  8.272e-10  3.403e-10   2.431 0.016123 *
## BMI                  3.607e+04  7.354e+03   4.904 2.19e-06 ***
## Incidents_HIV       -1.507e+05  5.109e+04  -2.950 0.003628 **
## BMI:Incidents_HIV    4.853e+03  1.987e+03   2.442 0.015643 *
## Schooling:BMI       -4.397e+03  8.811e+02  -4.990 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62620 on 169 degrees of freedom
## Multiple R-squared:  0.8501, Adjusted R-squared:  0.8421
## F-statistic: 106.5 on 9 and 169 DF,  p-value: < 2.2e-16
```

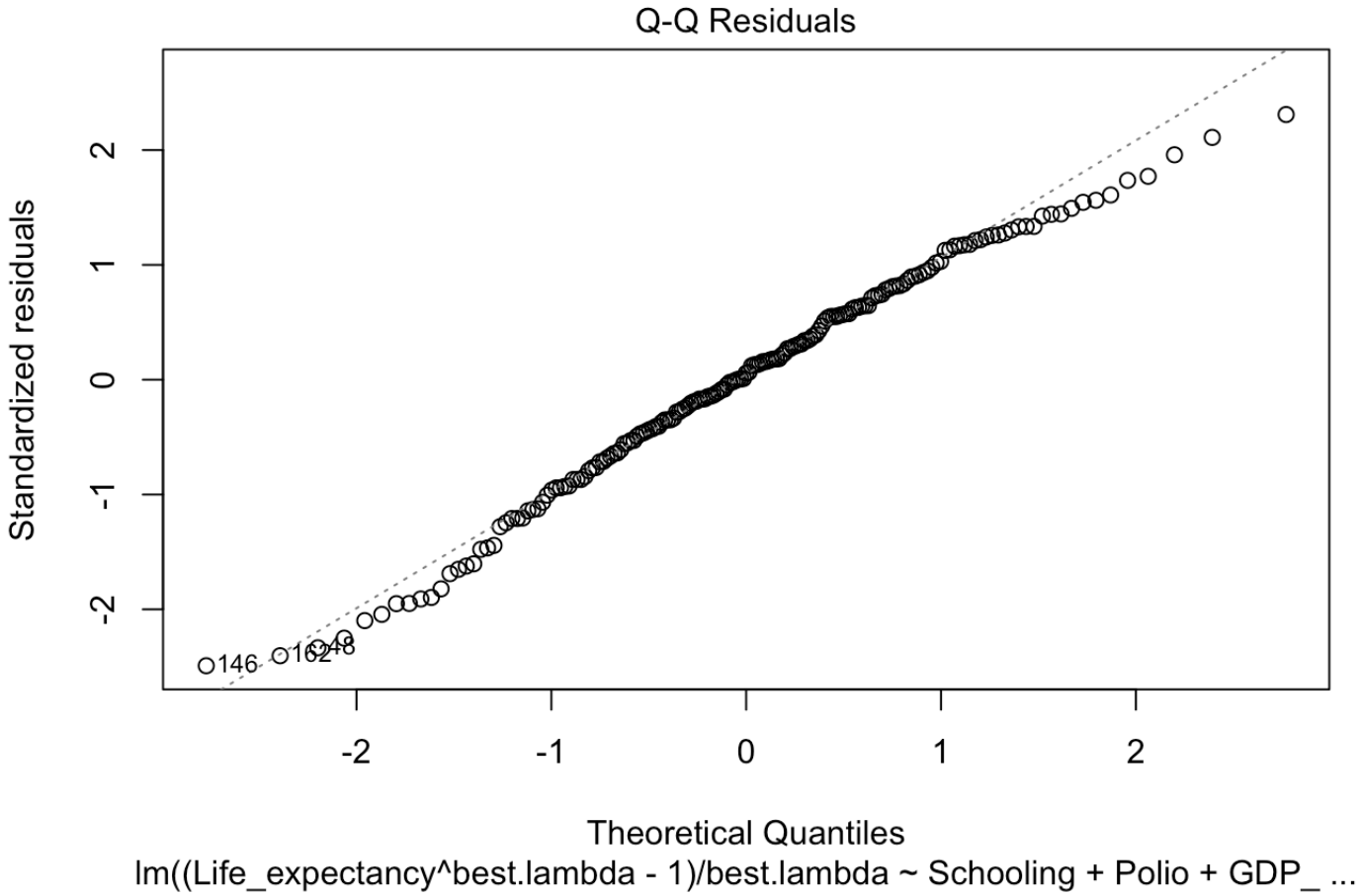
It appears that this box-cox still maintains all variables as being significant. Now lets test the normality assumption again:

Normality of the Residuals:

```
hist(residuals(boxcoxmodel), main = "Residual Histogram", xlab = "Residuals")
```




```
plot(boxcoxmodel, which = 2)
```




```
shapiro.test(residuals(boxcoxmodel))
```

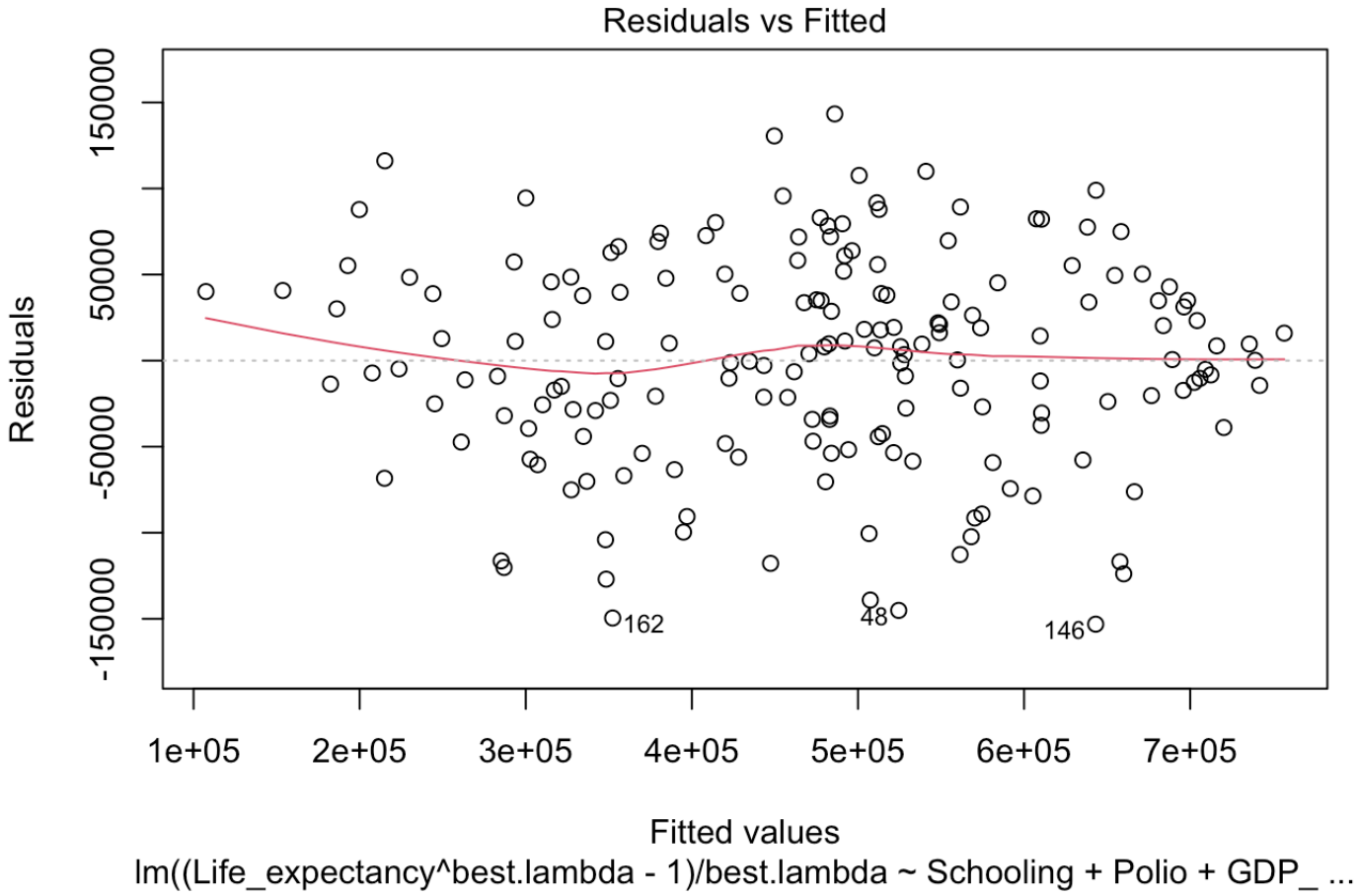
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(boxcoxmodel)
## W = 0.98846, p-value = 0.1535
```

$H_0: $ Residuals are normally distributed. $H_A: $ Residuals are not normally distributed.

According the Shapiro-Wilk the p-value is above 0.05, therefore we fail to reject the null hypothesis and can go ahead and conclude that the residuals are normally distributed. This is also backed up by our Q-Q normality plot and histogram.

Linearity:

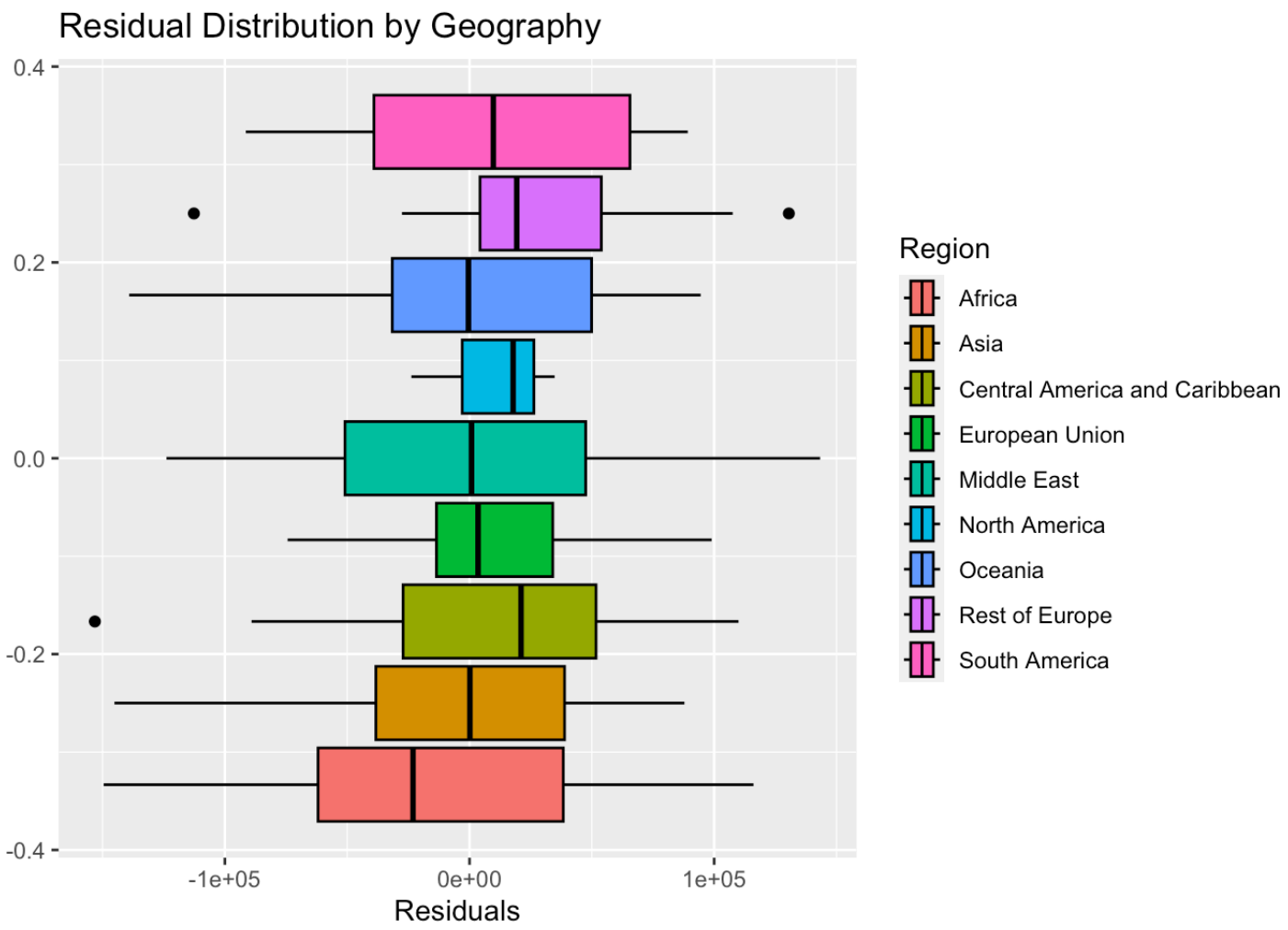
```
plot(boxcoxmodel, which = 1)
```



There appears to be no evident pattern, therefore it appears that the linearity assumption is met according the residual plot above.

Independence:

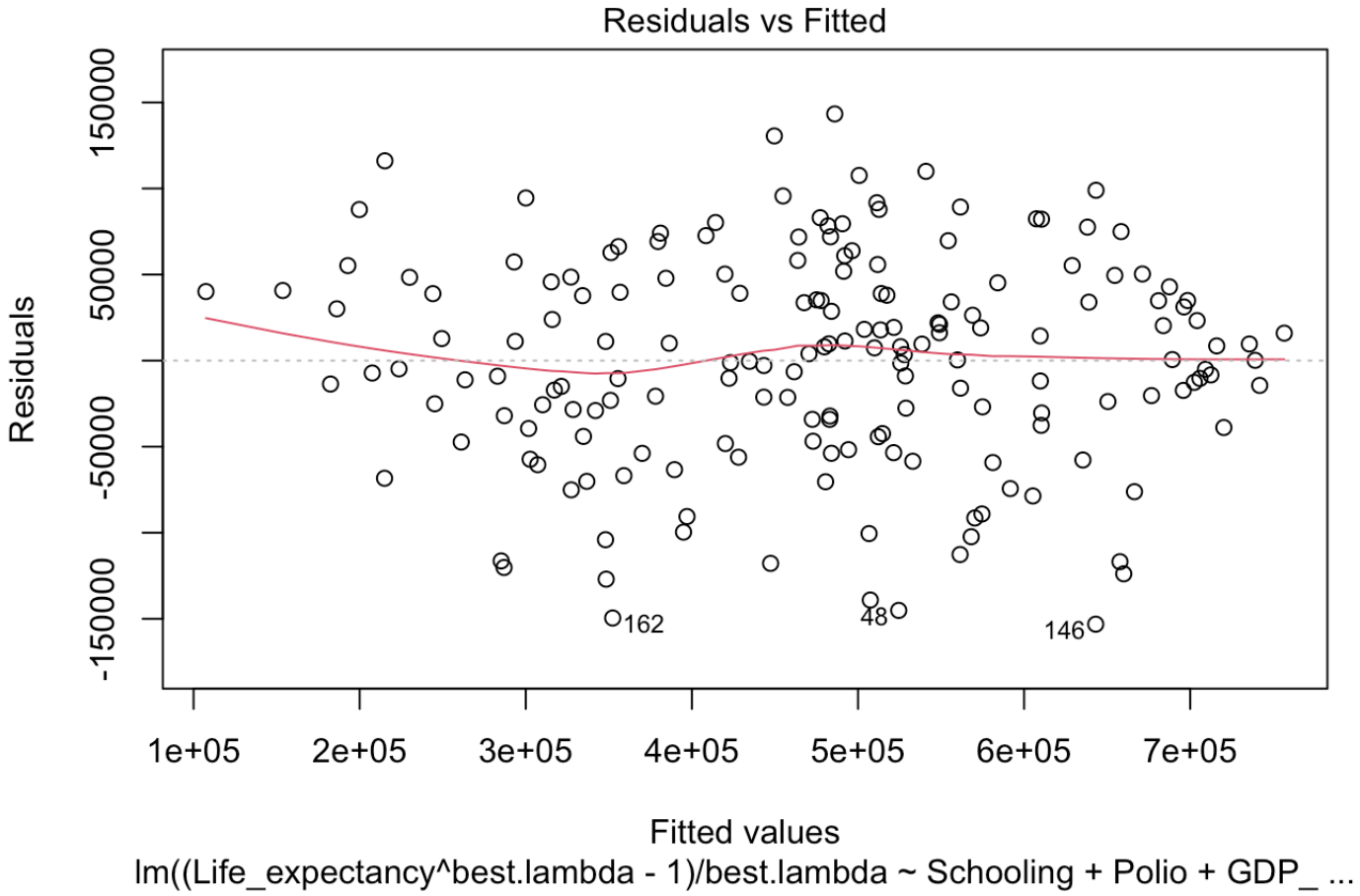
```
ggplot(data = data.frame(Residuals = residuals(boxcoxmodel), Region = life$Region), m
apping = aes(x = Residuals, col = "black", fill = Region)) +
  geom_boxplot(col = "black") + ggtitle("Residual Distribution by Geography")
```



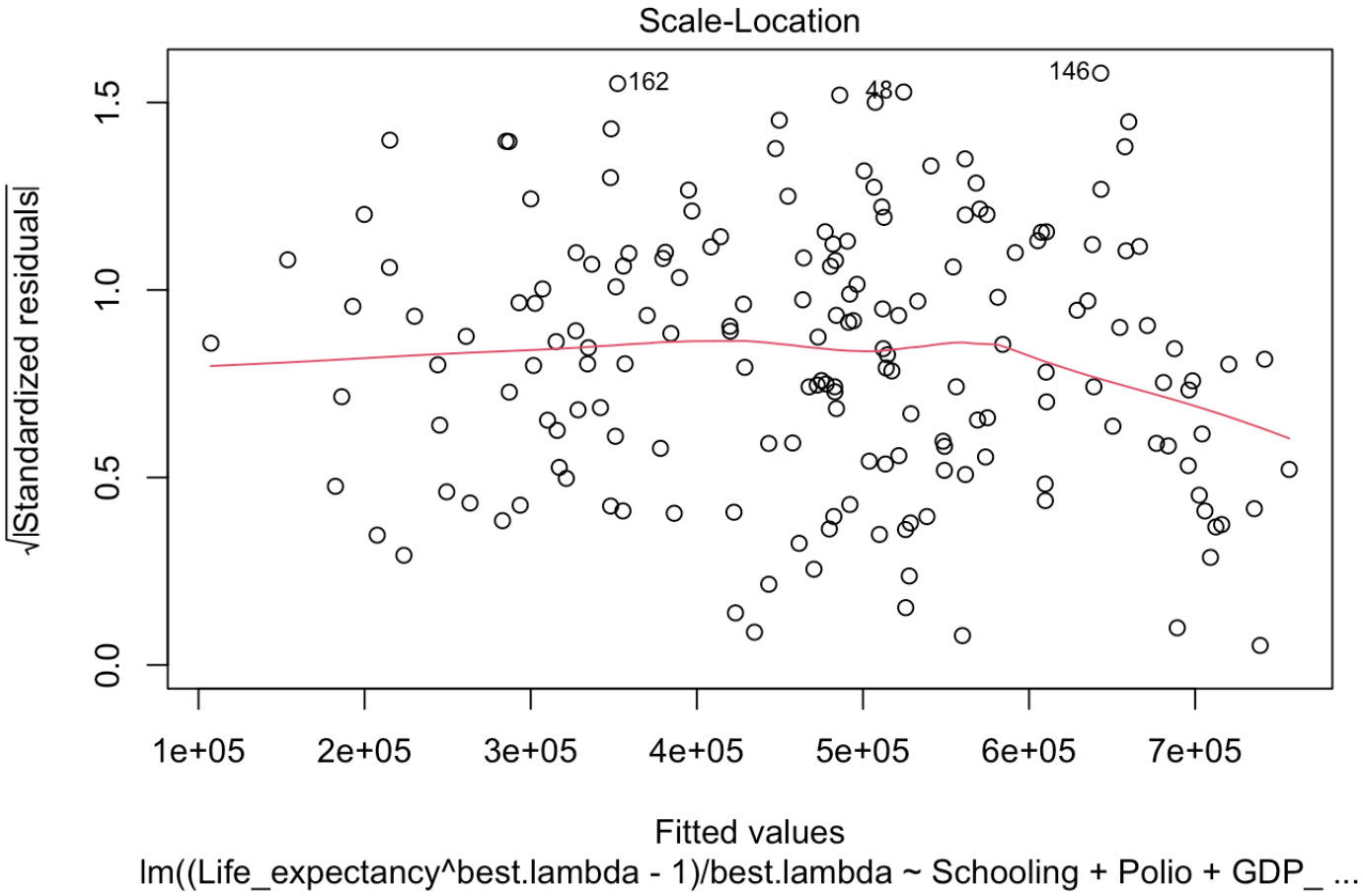
Independence is assumed by default, however in order to mitigate any dependencies, we ensured to filter for the year 2015 only in this training dataset. In addition, we also checked for clumping of residuals by geography in our boxplot above, All of these boxplots for each region overlap with eachother, showing there is no clumping going on by region.

Equal Variance:

```
plot(boxcoxmodel, which = 1)
```




```
plot(boxcoxmodel, which = 3)
```




```
bptest(boxcoxmodel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  boxcoxmodel
## BP = 9.3408, df = 9, p-value = 0.4064
```

$H_0: $ Homoscedacity holds. $H_A: $ Heteroscedacity holds.

It appears due to our Breusch-Pagan test, that the condition of equal variances holds, as the p-value is greater than 0.05. This is further confirmed by our residual plot and scale-location plots as it appears that the points are distributed equality above and below the line for all fitted values.

Multicollinearity:

We can check for multicollinearity in our current model, however it contains interaction/higher order terms, thus there will be high VIF values for it. However we can ignore this as we know this is due to those terms, and we can test the base model terms for any multicollinearity.

```
imcdiag(boxcoxmodel, method = "VIF")
```

```
##
## Call:
## imcdiag(mod = boxcoxmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                          VIF detection
## Schooling           219.9040         1
## Polio                 1.3682         0
## GDP_per_capita       44.1921         1
## I(GDP_per_capita^2) 171.1142         1
## I(GDP_per_capita^3)  63.0176         1
## BMI                  11.7918         1
## Incidents_HIV       311.4246         1
## BMI:Incidents_HIV   309.3098         1
## Schooling:BMI       285.3458         1
##
## Multicollinearity may be due to Schooling GDP_per_capita I(GDP_per_capita^2) I(GDP
_per_capita^3) BMI Incidents_HIV BMI:Incidents_HIV Schooling:BMI regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

We can also check for multicollinearity between the base terms, as our VIF test above included interaction, affected the VIF test negatively:

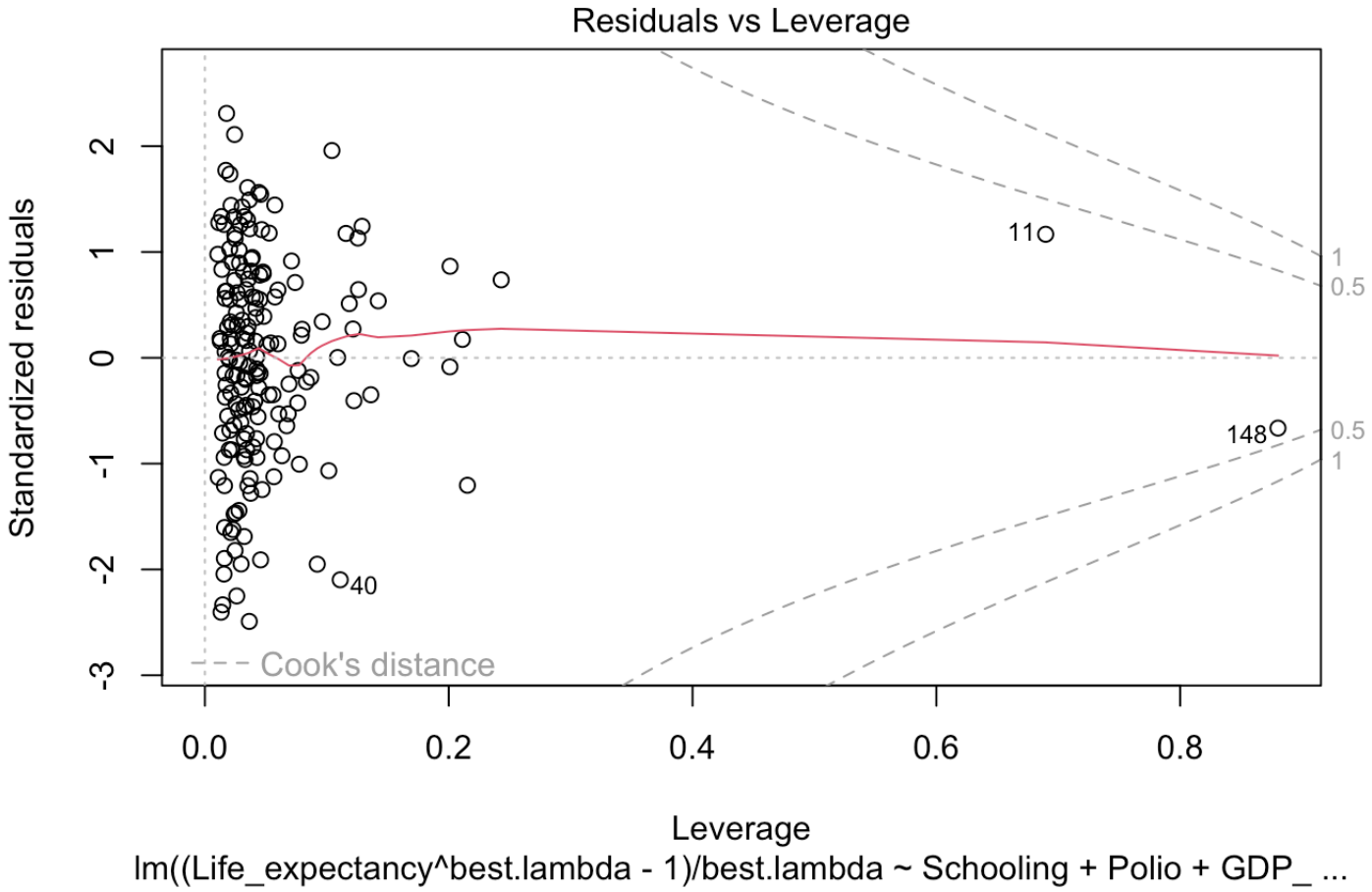
```
imcdiag(reducedmodel, method = "VIF")
```

```
##
## Call:
## imcdiag(mod = reducedmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                    VIF detection
## BMI             1.6106         0
## Polio           1.3420         0
## Incidents_HIV   1.0402         0
## GDP_per_capita 1.5959         0
## Schooling       2.6561         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

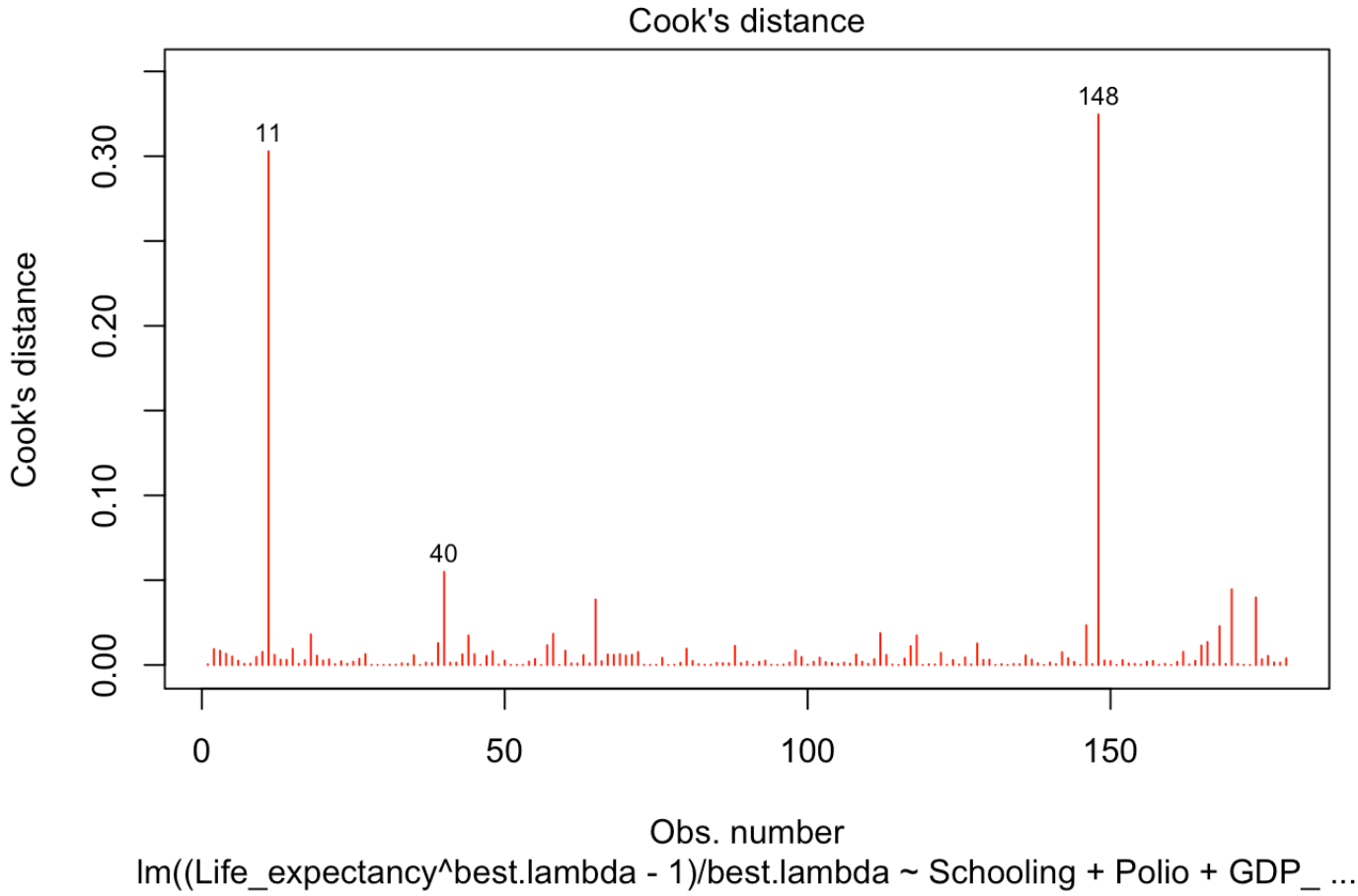
It appears that we can say that the model meets the multicollinearity assumption as, it appears there is no multicollinearity present between the base variables.

Outliers:

```
plot(boxcoxmodel, which = 5)
```



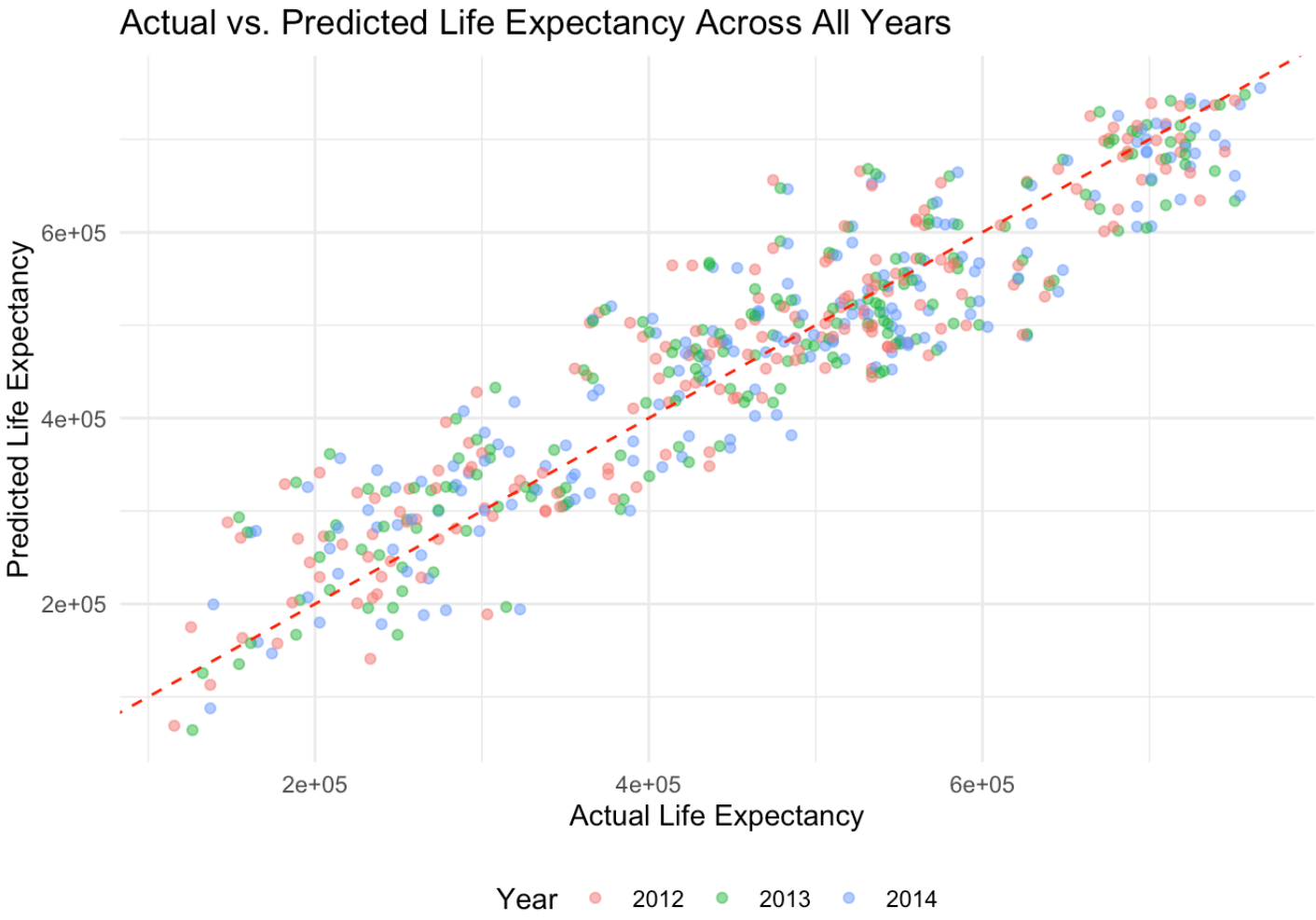

```
plot(boxcoxmodel, pch = 18, col = "red", which = c(4))
```



It appears, according to the cooks distance and residuals vs. leverage plot, there are no outliers, as all cooks distances are below 0.5.

Now, just to test how well our regression model does against new data, we can plot our predicted values against the actual values for the last 3 years of data, and it should follow the diagonal line if it is indeed an accurate model.

```
life.df$predicted_life_expectancy = predict(boxcoxmodel, newdata = life.df)

ggplot(life.df %>%
  filter(Year %in% c(2014, 2013, 2012)), aes(x = (Life_expectancy^best.lambda-1)/bes
t.lambda, y = predicted_life_expectancy)) +
  geom_point(aes(color = as.factor(Year)), alpha = 0.5) +
  scale_color_discrete(name = "Year") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Actual Life Expectancy", y = "Predicted Life Expectancy",
       title = "Actual vs. Predicted Life Expectancy Across All Years") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



As we can see, all of these points are very tightly clustered around the diagonal line, signifying an accurate and precise prediction. It also appears that the model is better when predicting for countries with a higher life expectancy, as those points seem to be more tightly clustered around the diagonal line, as opposed to ones with a lower life expectancy. This may be due to the fact that there may be more factors involved when considering regions with a lower life expectancy, than the ones included in our dataset.