# Linear Regression Analysis on Life Expectancy of Countries Worldwide in the Year 2015

Lucy Zeng, Chengzu Wu, Juanita Yang, Kalee Chan

June 17, 2024

## Table of Contents

**Part I: Introduction (Analysis Overview)**

**Part II: Statistical Methods**

**Part III: Results**

**Part IV: Discussions**

## (i) Introduction

The subject matter is particularly relevant in today's globalized society where people across the globe are constantly seeking ways to promote health and longevity. Life expectancy can vary greatly between countries across the world, as different countries have various conditions that influence the impacting factors. Understanding the determinants of life expectancy is essential to gaining further insights regarding the effectiveness of health policies, the influence of socioeconomic conditions, and the impact of lifestyle factors. Therefore, the subject matter is not only of personal interest but also an important matter to public health practitioners and policymakers aimed at improving global health conditions.

Many qualified researchers have delved into the topic of life expectancy, analyzing historical trends, and concluding with factors that may affect life span. In the research done by Dattani (et al., 2023), it was noted that the high child mortality rate was a key limiting factor. However, this issue has become less severe in recent years as a result of the advancements in medicine, public health, and living standards. Beyond healthcare, an interesting point written in the works by Mackenbach (et al., 2019) was that an individual's education level is strongly influential to one's longevity. This is an interesting take, as one may not have otherwise assumed a strong correlation. One of the most comprehensive research was published by the NIH by Galvani-Townsend (et al., 2022), categorizing contributing factors into 5 major categories that determine life expectancy: economic stability, education, health & health care, neighbourhood & built environment, and social & community context.

Although the topic of life expectancy is currently a popular area of research, we look to contribute to this field by further refining the existing model in a linear framework. Despite many pieces of research including a plethora of predicting variables, we expect there to be problematic properties of the existing models. For example, when adult mortality is taken into account, other factors such as death rates of many communicative diseases may logically be largely correlated, therefore violating the assumption of independence between predictors. This motivates our research question: *In a linear framework, what are the most significant determinants of life expectancy and do they suffice to reflect the real observations across countries?* We will proceed by utilizing linear regression methods on existing data which will be able to provide an insight on how each of the selected predictors contribute to life expectancy, and which ones are the most influential. The life expectancy measured in years will be the selected response variable, while the predicting variable all belong to the five main categories of influential factors of the response variable.

## (ii) Methods

### 2.1 Overview

This linear regression analysis was performed on independently sampled data of 183 countries' life expectancy and its potentially influential factors in 2015. The sample size of 183 was extracted from a compiled dataset on Kaggle sourced from the United Nations (Russell, et al., 2018). The data was further isolated for the most recent year, 2015.

To identify an optimal linear regression model, we began by preprocessing to filter the dataset, removing any missing or non-applicable values. We take this as Model 1 so that its predictors consist of all factors in the 2015 data. We want to test the performance of Model 1 against a reduced model we will refer to as Model 2. Removing predictors of infant deaths, Polio, Thinness, HIV/AIDS, Measles, and Hepatitis B from Model 1 leaves us with Model 2 incorporates a subset of the predictors: adult mortality, BMI, Diphtheria immunization coverage, and number of years of schooling. In short, Model 2 was developed using the stepwise method of backwards elimination, where we start with all the impacting predictors and reduce down to the most relevant. This process will be further improved and elaborated on in detail so that we can test our model's performance with the existing literature's model, Model 1.

### 2.2 Model Diagnostics & Assumption Checks

The independently sampled observations were split randomly to validate our model estimates in a 60/40 ratio: 60% allocated as a training dataset and the remaining 40% to evaluate the performance of each model. We will perform all model building and diagnostics using only the training dataset, then fit the 'best' model in the testing dataset.

Before attempting to fit models, an exploratory analysis was performed, where a plot diagnostic was run to ensure the data was appropriate to be addressed by linear regression. This included analyzing the pairwise scatterplot graphs of each predictor with the response variable to identify if the relationship is linear or nonlinear. Histograms were also used to observe whether the data exhibits a normal distribution. Moreover, we assessed the plots of Residual vs. Predictor, Residual vs. Fitted, and Normal Q-Q to check if any linear regression assumptions were violated. Desirable residual plots would display randomness and lack of apparent cluster of points, which would be a good indicator that independence (uncorrelated errors) and homoscedasticity (constant variance) assumptions were held. The Normal Q-Q Plot serves to show the relationship between the model's residuals' quantiles and the quantiles of a Normal distribution. Since it matches z-scores of the residuals to critical values of the standard Normal distribution, it is useful for assessing the normality of error assumption where a satisfactory plot would appear as a straight diagonal line with a positive slope and minimal deviations at the ends.

If there are any violations of constant variance, we can attempt to employ methods of variance stabilizing transformations, such as the frequently used square root and natural logarithm. The scatter and residual plots, as well as the Box-Cox method, can help to decide on simple yet adequate transformations to employ. With a reasonable function applied to the response, the transformed variable may have a more constant variance.

### 2.3 Building Model 2

Moving forward, we can begin constructing our reduced model, Model 2. To determine which predictors to eliminate, we employed a backward elimination method, retaining only the relevant predictors with p-value < 0.05 (significance indication). To ensure accuracy, we also checked all pairwise correlations, aiming for a lower correlation among the pairwise predictors while each predictor should have a high linear correlation with the response variable of life expectancy. This multicollinearity can be assessed using a one-to-one collinearity scatterplot matrix, as well as finding the Variance Inflation Factor (VIF) and removing predictors with VIF > 5.

**2.4 Problematic Observations (Model Adjustments)**

Once we have a valid model, we check for problematic observations to ensure the fitted model is indeed doing a reasonable job fitting the data. That is, we look for any leverage points, invalid outliers, or influential observations that must be dealt with. The rule for identifying leverage points in our multiple linear regression is if the $h_{ii}$ *is greater than twice the average($h_{ii}$).* A leverage point is ‘bad’ if it is too distant from the overall centre of mass of all the predictors so we remove the invalid point that doesn’t agree with the regression (like an outlier). As the dataset is moderately sized, if the standardized residual of a point falls outside the interval (-2, 2), we can also eliminate them as outliers. To examine the influence of a single observation, we can evaluate if Cook’s Distance is greater than the 50th percentile of $F(p,\ n - p)$.

**2.5 Model Selection**

Our model selection was guided by the principle of *Occam’s Razor* and the four criteria: preference for large $R^2_{adj}$ value, small AIC (Akaike’s Information Criterion) and corrected AIC, and small BIC (Bayesian Information Criterion). The preferred model would have a relatively large adjusted $R^2_{adj}$ value paired with fewer predictors. Meaning, we ensure that each predictor has a significant enough improvement on the goodness-of-fit measure to justify its inclusion. By ensuring that corresponding partial F-test statistics > 0.05, we can deduce that the included predictors explain a reasonable amount of residual variation. Requiring a small AIC and BIC is to ensure that the log-likelihood function contributes enough to outweigh the penalty term, meriting a more complex model. Using the aforementioned criteria, we can observe whether Model 1 or Model 2 is preferred.

**2.6 Model Validation (Final Step)**

Once we confirm that all assumptions are met in our preferred model, we will re-perform the diagnostics and analysis using the testing dataset, looking out for large violations of assumptions as this would indicate inconsistencies in predictions. We can evaluate the model’s performance by comparing the results and properties (mean, median, variances) of the test dataset against the training. To validate our model, we want to observe similar characteristics.

In particular, we look for:

- Minimal differences in the estimated regression coefficients, especially not off by larger than the standard error or each coefficient in the training data.
- The same predictors appear significant to confirm we did not overfit data.
- Similar adjusted goodness-to-fit $R^2_{adj}$ value (similarly good at explaining variation)
- Common issues due to problematic observations (model is influenced similarly)
- Similar multicollinearity observed, with similar severity of impact (test dataset should not escalate the observation into a problematic situation.

If the model can not be validated, it may have been over-fit, under-fit, or suffered different impacts of influential points. In that case, we will discuss these limitations to our model.

### (iii) Results

The following two tables show that the numerical summaries of the training and testing datasets show similar results, which supports our model validation.

**Numerical Summary of Variables in Training Dataset**

| | Life expectancy | Adult Mortality | BMI | Diphtheria | Schooling |
|---|---|---|---|---|---|
| MIN | 46.11 | 18.75 | 12.88 | 26.81 | 4.019 |
| 1st Qu. | 59.45 | 100.83 | 18.08 | 66.95 | 8.652 |
| Median | 69.60 | 159.12 | 37.91 | 83.03 | 11.275 |
| Mean | 67.58 | 173.34 | 35.15 | 78.42 | 11.289 |
| 3rd Qu. | 74.33 | 231.53 | 50.40 | 94.94 | 13.717 |
| Max | 82.52 | 462.38 | 62.94 | 98.94 | 20.038 |

**Numerical Summary of Variables in Testing Dataset**

| | Life expectancy | Adult Mortality | BMI | Diphtheria | Schooling |
|---|---|---|---|---|---|
| MIN | 48.78 | 54.12 | 14.79 | 48.44 | 7.125 |
| 1st Qu. | 65.28 | 115.06 | 24.81 | 76.62 | 10.722 |
| Median | 71.73 | 139.94 | 43.84 | 87.38 | 12.231 |
| Mean | 69.04 | 175.85 | 39.17 | 84.37 | 12.340 |
| 3rd Qu. | 73.81 | 217.62 | 51.16 | 94.62 | 13.628 |
| Max | 81.22 | 550.06 | 69.43 | 98.75 | 16.600 |

*Table 1: Numerical Summary of Variables*

The histograms displayed in Figure 1 and Figure 2 show that the training and testing datasets exhibit similar skewness in terms of frequency distribution.

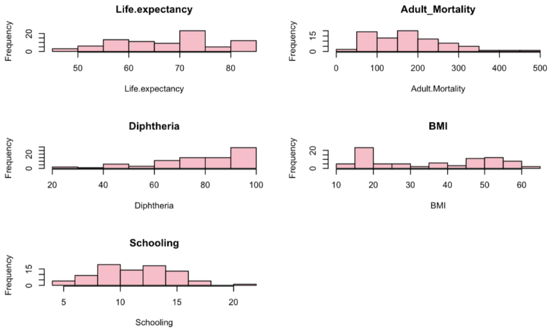


*Figure 1: Frequency Distribution of Training Dataset*

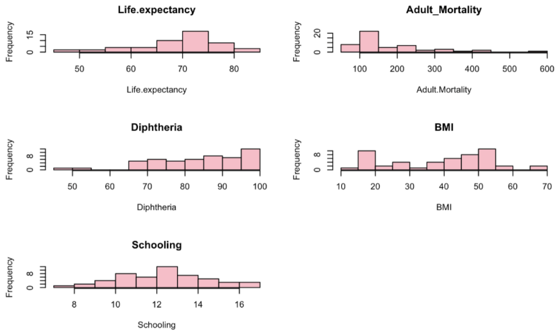


*Figure 2: Frequency Distribution of Testing Dataset*

From Figure 3 below, it can be observed that the Residuals vs Fitted Plot is patternless, and the Normal Q-Q Plot lies well on the diagonal line. This suggests that our model reasonably satisfies the linear regression analysis assumptions of predictor independence, homoscedasticity, and normality of errors.

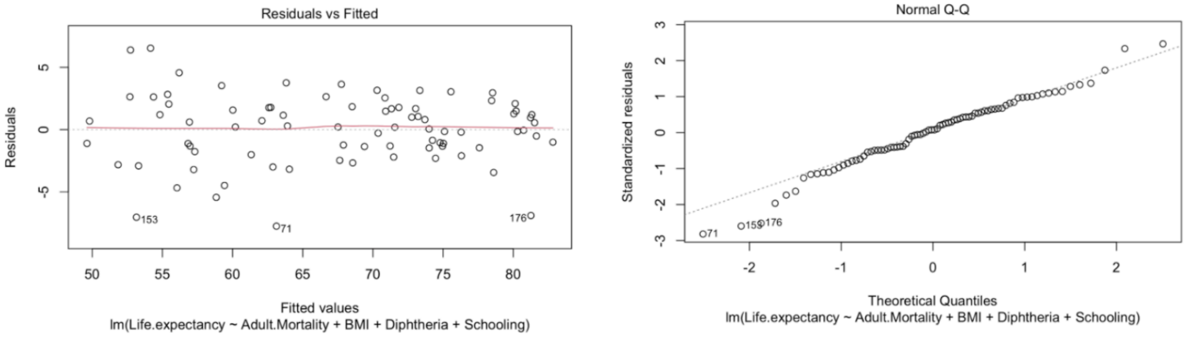


*Figure 3: Residuals vs Fitted and Normal Q-Q*

Additional Conditions 1 and 2 are checked in Figure 4 below:

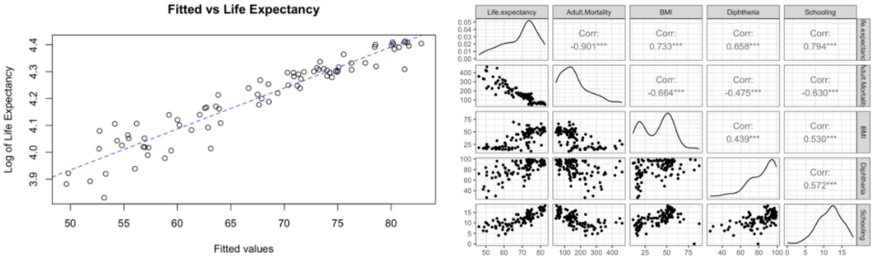


*Figure 4: Fitted vs Life Expectancy and Scatterplot Matrix*

The Fitted vs. Life Expectancy appears to respect linear relationships, as desired. However, the correlation between some of the predictors is high, and there appears to be a linear relationship between predictors. Thus, we perform the Box-Cox transformation to reduce the correlation in Figure 5, in which the scatterplot patterns between predictors appear to be random, as desired.

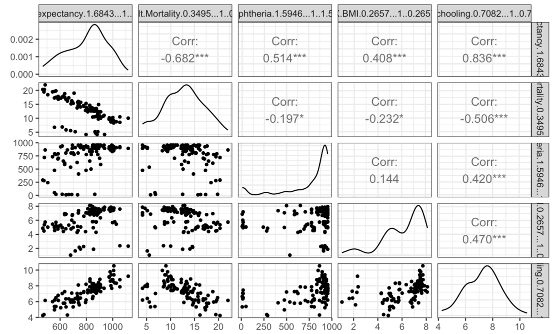


*Figure 5: Transformation of Scatterplot Matrix*

By using backwards elimination to find an appropriate subset of predictors for Model 2, it resulted in a similar $R^2_{adj}$ value compared to Model 1. As we only have a few outliers and no influence on the result (*Appendix 4)*, there was no need to remove them.

In reference to Table 2, the VIF of our predictors from Model 2 are less than 5, indicating the independence of each predictor. As preferred, Model 2 also has a smaller AIC and BIC (Table 3). The higher F-statistic in Model 2 also indicates the model provides a better fit to the data.

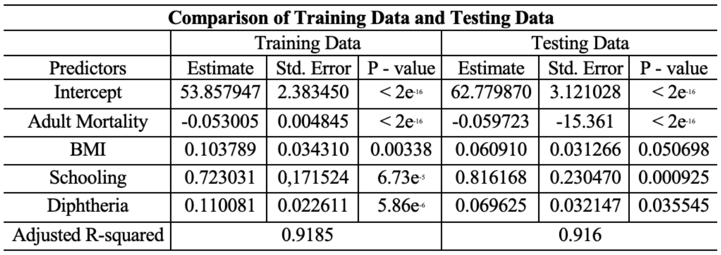
| **Predictors** | Adult Mortality | BMI | Schooling | Diphtheria |
|---|---|---|---|---|
| **VIF** | 2.00979 | 3.226809 | 3.476470 | 1.769094 |

*Table 2: Predictors' Variance Inflation Factor*

| | R-squared | Adjusted R-squared | F-statistic | AIC | AIC_c | BIC |
|---|---|---|---|---|---|---|
| Model 1 | 0.9298 | 0.9188 | 84.33 | 176.4764647 | 180.8708309 | 209.3570956 |
| Model 2 | 0.9225 | 0.9185 | 229.2 | 172.4354749 | 173.5263840 | 190.8757904 |

*Table 3: Comparison of the two models*

Lastly, we further validated our model using the testing dataset by comparing its estimated coefficients and p-values with the training set, noting a similar $R^2_{adj}$ . As depicted in the following Table 4, it is evident that all predictors' estimated coefficients are within their standard errors and are significant in both datasets. Moreover, we checked the Residuals vs. Fitted Plot and the Normal Q-Q plot (Appendix 2).



**Comparison of Training Data and Testing Data**

| | Training Data | | | Testing Data | | |
|---|---|---|---|---|---|---|
| Predictors | Estimate | Std. Error | P - value | Estimate | Std. Error | P - value |
| Intercept | 53.857947 | 2.383450 | $< 2e^{-16}$ | 62.779870 | 3.121028 | $< 2e^{-16}$ |
| Adult Mortality | -0.053005 | 0.004845 | $< 2e^{-16}$ | -0.059723 | -15.361 | $< 2e^{-16}$ |
| BMI | 0.103789 | 0.034310 | 0.00338 | 0.060910 | 0.031266 | 0.050698 |
| Schooling | 0.723031 | 0,171524 | $6.73e^{-5}$ | 0.816168 | 0.230470 | 0.000925 |
| Diphtheria | 0.110081 | 0.022611 | $5.86e^{-6}$ | 0.069625 | 0.032147 | 0.035545 |
| Adjusted R-squared | 0.9185 | | | 0.916 | | |

*Table 4: Comparison of Training Data and Testing Data*

After a thorough examination of the testing dataset, Model 2 is validated.

**(iv) Discussion**

Our analysis was aimed at constructing an accurate and parsimonious linear regression model for predicting life expectancy, with fewer determinants to predict the real observation. The final model retained the most significant and appropriate predictors: Adult Mortality, BMI, Diphtheria, and Schooling. This model provides worthy insights into how the key factors are correlated with population longevity across different countries. Other than Adult Mortality being negatively correlated, BMI, Diphtheria, and Schooling are all positively related. We found that the year of schooling has emerged as the most influential positive predictor of life expectancy among the four variables. Each additional year of schooling is linked to a large increase in life expectancy, suggesting the substantial benefits of education on health and longevity. Likewise, both BMI and Diphtheria immunization (coverage among 1-year-olds) show a highly significant positive relationship with life expectancy. A higher BMI is associated with a slight increase in life expectancy, while higher immunization rates against diphtheria are strongly linked to increased life expectancy, indicating that good nutritional status and effective vaccination contribute to longer lifespans. Additionally, higher adult mortality rates are strongly negatively associated with life expectancy. Given the context of the real-world relationship, it is not surprising that a higher mortality rate would correlate to a lower lifespan.

Despite our linear regression model providing a robust framework for understanding the most significant determinants of life expectancy, several limitations must be addressed. For starters, real-world relationships may be much more complex and involve non-linear relationships that our model may fail to capture. For instance, if a country is faced with heavy financial constraints, they may not even be able to afford living necessities like food, much less Schooling. Therefore, although Schooling may appear to be the factor largely correlated, in reality, there are many other factors influencing education access. Looking at the Residual Multicollinearity, a small linear correlation still exists between schooling and the other three predictors. Although the VIF for Schooling remains within 5 which is acceptable, it is the highest among all four predictors, suggesting a certain degree of multicollinearity, making it harder to isolate. To improve this flaw, an exploration of the factors contributing to Schooling is necessary for future research, which can help lower the VIF and correlation with other predictors. In addition, there is a discrepancy in p-values for BMI: The testing dataset yields a greater p-value than the training dataset. This may be due to the dataset that we selected having a small number of observations, suggesting that our model's findings will be better with larger samples.

## Reference List

Academic Papers:

Saloni Dattani, Lucas Rodés-Guirao, Hannah Ritchie, Esteban Ortiz-Ospina and Max Roser (2023) - "Life Expectancy" Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/life-expectancy'

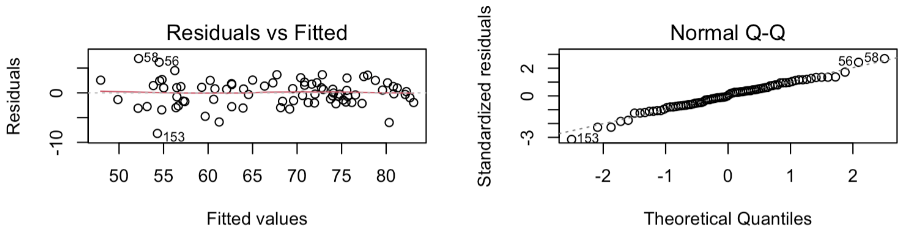
Mackenbach, J. P., Valverde, J. R., Bopp, M., Brønnum-Hansen, H., Deboosere, P., Kalediene, R., et al. (2019). Determinants of inequalities in life expectancy: An international comparative study of eight risk factors. *The Lancet Public Health*, *4*(10). doi:10.1016/s2468-2667(19)30147-1

Galvani-Townsend, S., Martinez, I., & Pandey, A. (2022). Is life expectancy higher in countries and territories with publicly funded health care? Global Analysis of Health Care Access and the social determinants of health. *Journal of Global Health*, *12*. doi:10.7189/jogh.12.04091

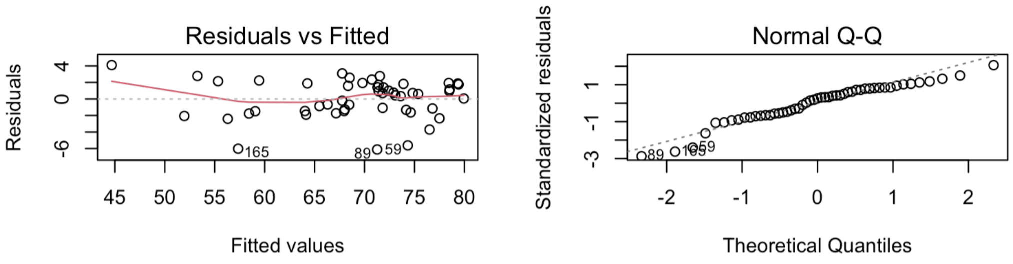
Data:

Russell, D., & Wang, D. (2018, February 10). Life expectancy (WHO). (K. Rajarshi, Ed.) *Kaggle*. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. Accessed 15 June 2024

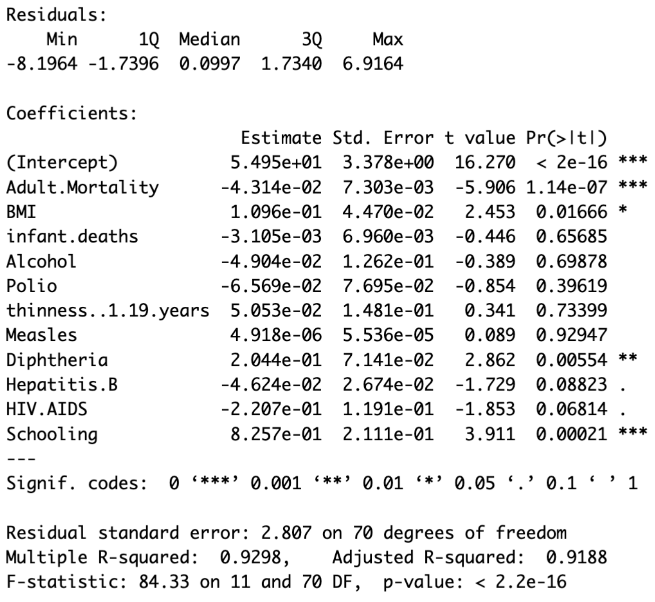
## Appendix



*Appendix 1: Figure for Model 1 training set (Residuals vs Fitted, Normal Q-Q)*



*Appendix 2: Model 2 testing set (Residuals vs Fitted, Normal Q-Q)*

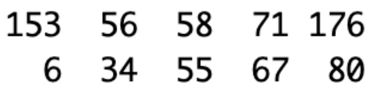


```
Residuals:
    Min      1Q  Median      3Q     Max 
-8.1964 -1.7396  0.0997  1.7340  6.9164 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           5.495e+01  3.378e+00  16.270  < 2e-16 ***
Adult.Mortality      -4.314e-02  7.303e-03  -5.906 1.14e-07 ***
BMI                   1.096e-01  4.470e-02   2.453  0.01666 *  
infant.deaths        -3.105e-03  6.960e-03  -0.446  0.65685    
Alcohol              -4.904e-02  1.262e-01  -0.389  0.69878    
Polio                -6.569e-02  7.695e-02  -0.854  0.39619    
thinness..1.19.years  5.053e-02  1.481e-01   0.341  0.73399    
Measles               4.918e-06  5.536e-05   0.089  0.92947    
Diphtheria            2.044e-01  7.141e-02   2.862  0.00554 ** 
Hepatitis.B          -4.624e-02  2.674e-02  -1.729  0.08823 .  
HIV.AIDS             -2.207e-01  1.191e-01  -1.853  0.06814 .  
Schooling             8.257e-01  2.111e-01   3.911  0.00021 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.807 on 70 degrees of freedom
Multiple R-squared:  0.9298,	Adjusted R-squared:  0.9188 
F-statistic: 84.33 on 11 and 70 DF,  p-value: < 2.2e-16
```

*Appendix 3: Figure for Model 1*



```
153  56  58  71 176 
  6  34  55  67  80 
```

*Appendix 4: Outliers for Model 2 trainset*