Life Expectancy*

Karmen Zhu, Edan Wong, Saanvi Prasanth

November 12, 2024

1 Introduction

2 Data

2.1 Raw Data

The data used in this paper is access in from Open Data Toronto and the particular data set used was the Daily Shelter & Overnight Service Occupancy & Capacity ((**opendatatoronto?**)). To analysis the data and creating graphs using the data, following package that was build in the (R program R Core Team (2023)) was used: tidyverse (Wickham et al. (2019)), dplyr (Wickham et al. (2023)), lubridate ((**lubridate?**)), and ggplot2 (Wickham (2016)). We clean the column names, separate the date into `year` and `month`, and create a standardized date column for monthly aggregation. The cleaned data is then saved for further analysis.

2.2 Response variable(Life Expectancy)

	LifeExpectancy	GDP	Polio	Diphtheria
LifeExpectancy	1.0000000	0.4619693	0.4527720	0.4658321
GDP	0.4619693	1.0000000	0.2098406	0.1983948
Polio	0.4527720	0.2098406	1.0000000	0.6801445
Diphtheria	0.4658321	0.1983948	0.6801445	1.0000000
IncomeComposition	0.7324831	0.4615884	0.3908915	0.4149200

	IncomeComposition
LifeExpectancy	0.7324831
GDP	0.4615884

*Code and data are available at: _____

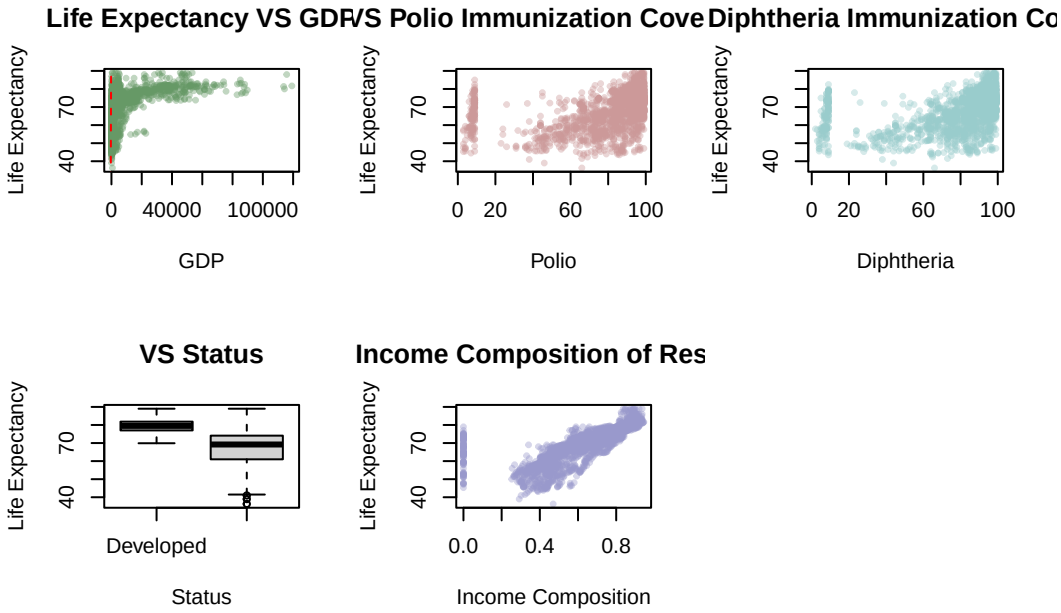

Figure 1: Visualizing the Relationship Between Life Expectancy and Its Predictors

```
Polio                   0.3908915
Diphtheria              0.4149200
IncomeComposition       1.0000000
```

The correlation matrix shows the high correlation between `Polio` and `Diphtheria`, which means that `Polio` and `Diphtheria` is linearly dependent, and including both predictor would lead to multicollinearity. To address this issue, we decided to drop one of these two predictors from the model, given that both variables capture similar information related to immunization coverage.

```
Call:
lm(formula = LifeExpectancy ~ GDP + Polio + Diphtheria + Status +
    IncomeComposition + BMI, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-28.1590  -2.9574   0.3713   3.0937  22.0416

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.030e+01  1.028e+00  48.950  < 2e-16 ***
```

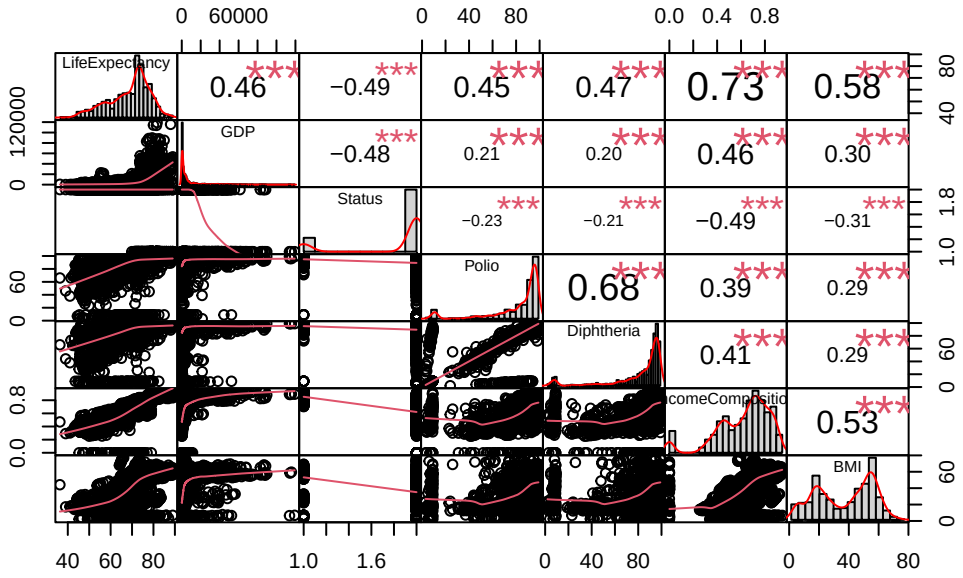

Figure 2: Correlation Plot

```
GDP                 6.608e-05  1.001e-05   6.599 5.13e-11 ***
Polio               3.991e-02  7.289e-03   5.476 4.84e-08 ***
Diphtheria          4.784e-02  7.265e-03   6.585 5.62e-11 ***
Status             -2.858e+00  3.831e-01  -7.459 1.23e-13 ***
IncomeComposition   1.936e+01  7.940e-01  24.387  < 2e-16 ***
BMI                 1.118e-01  7.304e-03  15.301  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.822 on 2294 degrees of freedom
Multiple R-squared:  0.6413,    Adjusted R-squared:  0.6404
F-statistic: 683.5 on 6 and 2294 DF,  p-value: < 2.2e-16
```

GDP: Shows fan-shaped pattern of the distribution instead of null pattern and the data points scattered around zero, which violates constant variance. The distribution of data does not have clusters, so the assumption of uncorrelated error holds. The scatter plots does not have a systematic pattern, which satisfies the assumption of linearity.

Polio&Diphtheria: Shows fan-shaped pattern of the distribution instead of null pattern and the data points scattered around zero, which violates constant variance. The distribution of data does not have clusters, so the assumption of uncorrelated error holds. The scatter plots does not have a systematic pattern, which satisfies the assumption of linearity.

Income Composition: No fan-shaped pattern, the data points are gathered together, so it does not violate homoscedasticity assumption, as well as the assumption of uncorrelated error

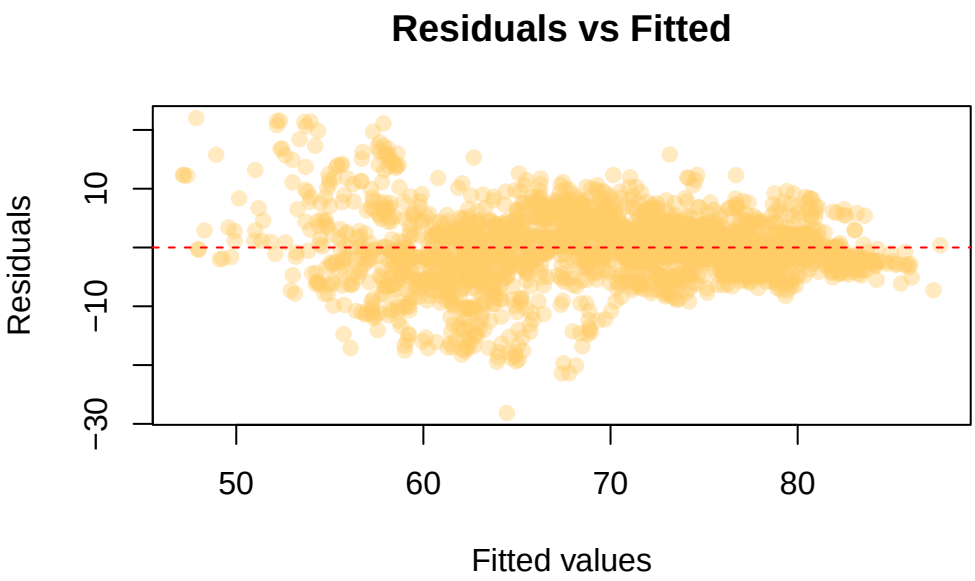

Figure 3: Residuals vs Fitted

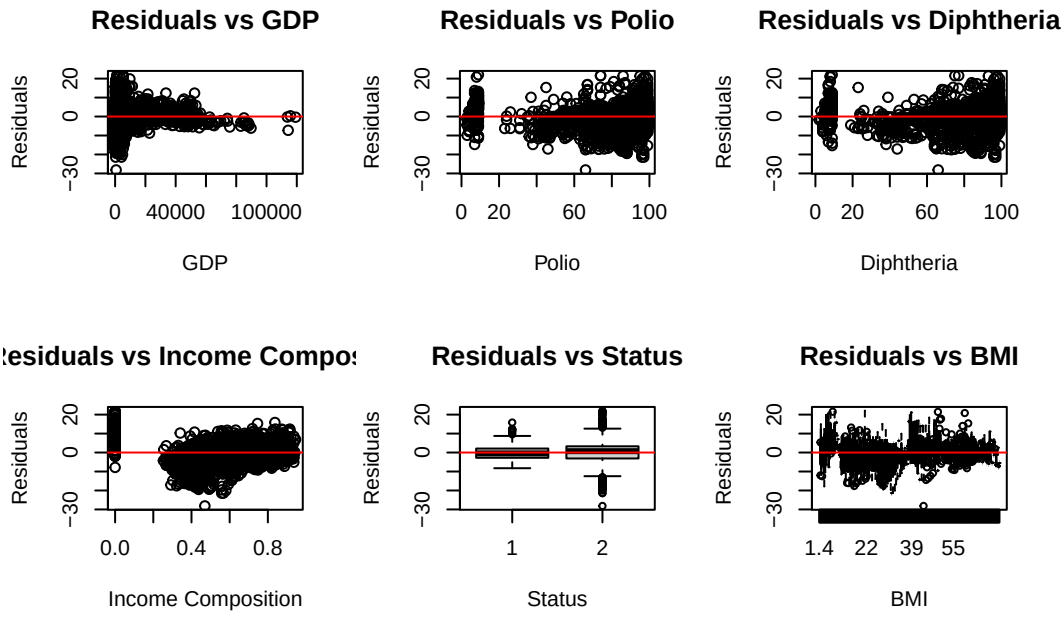

Figure 4: Residuals versus Each Predictor

holds because the distribution of data does not have clusters. However, a linear pattern of the scatter plot is presented, which violates the assumption of linearity.

Status: the skewed boxplot in Residual vs Status indicate that residuals is not normalized, this violates normality assumption, which could affect inference and p-values. The residuals across the boxplots for each status (`1` stands for developed and `2` stands for developing) is widely spread, which suggests a violation of constant variance. The meidan line of residuals for each `Status` category should be close to 0, the Linearity holds for Status

BMI: No fan-shaped pattern, the data points are gathered together, so it does not violate homoscedasticity assumption, as well as linearity due to no systematic pattern shown While there are two clusters presented in the scatter plot, which causes a violation of uncorrelated error.

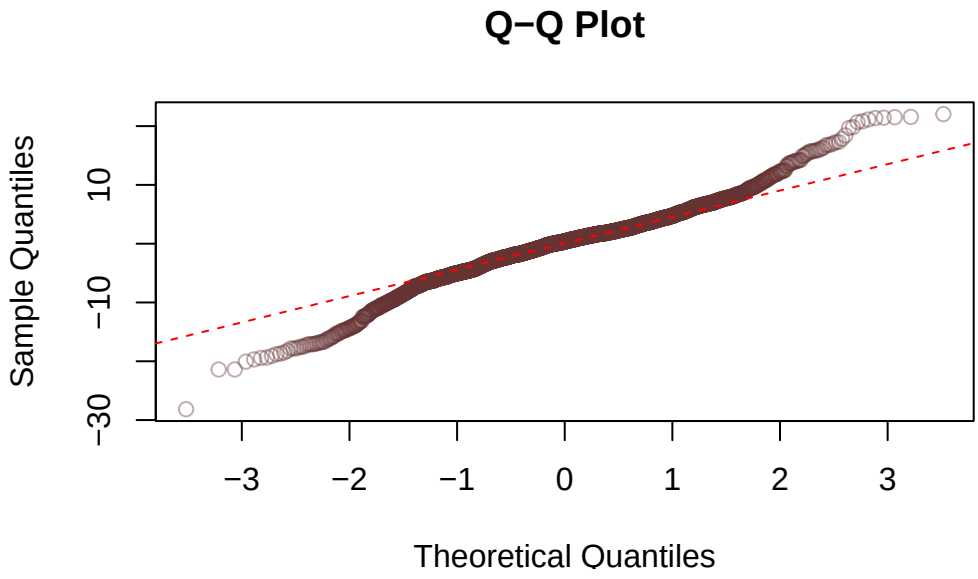

Figure 5: Q-Q Plot of Residuals

The QQ-plot of residuals are close to the 45-degree line with deviations at the tails, which means the assumption of normality is violated.

The distribution of residuals is normally distributed, meaning that this satisfies normality assumption.

```
bcPower Transformation to Normality
                           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
data_clean$LifeExpectancy    2.9605        2.96       2.6556       3.2655

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
```

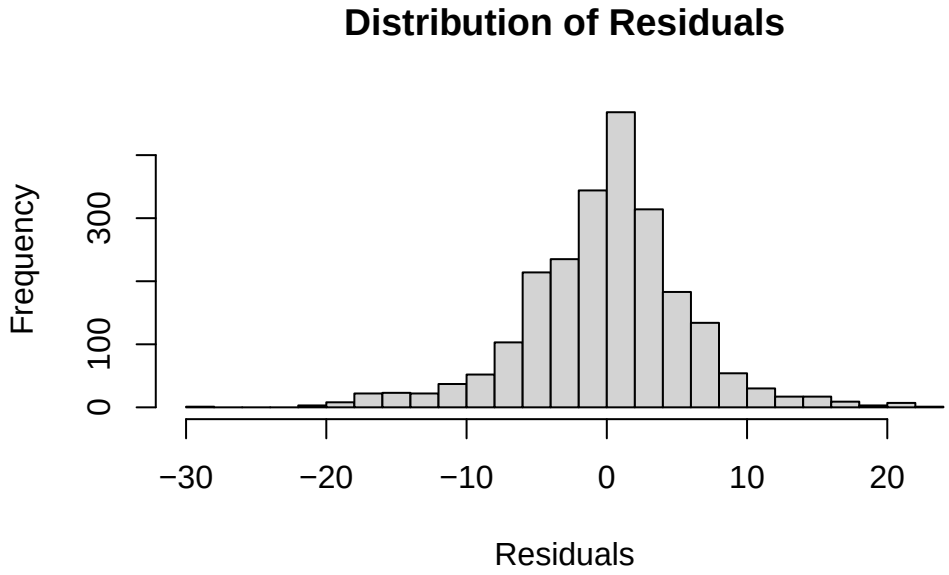

Figure 6: Distribution of Response Value

```
                          LRT df       pval
LR test, lambda = (0) 397.331  1 < 2.22e-16

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 168.9374  1 < 2.22e-16


bcPower Transformations to Multinormality
                  Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
GDP                  0.0696        0.07       0.0510       0.0881
Polio                3.9222        3.92       3.7708       4.0735
Diphtheria           3.9662        3.97       3.8131       4.1194
IncomeComposition    1.4341        1.43       1.2958       1.5723
BMI                  1.0873        1.09       1.0168       1.1577

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                                   LRT df       pval
LR test, lambda = (0 0 0 0 0) 11535.63  5 < 2.22e-16

Likelihood ratio test that no transformations are needed
                                   LRT df       pval
LR test, lambda = (1 1 1 1 1) 10127.04  5 < 2.22e-16
```

```
Call:
lm(formula = LifeExpectancy_transformed ~ GDP_transformed + Polio_transformed +
    Diphtheria_transformed + Status + IncomeComposition_transformed +
    BMI_transformed, data = data_clean)

Residuals:
   Min     1Q Median     3Q    Max
-53571  -9583    705   9541  60065

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.762e+05  3.711e+03  47.489  < 2e-16 ***
GDP_transformed               -5.271e+02  1.702e+02  -3.097  0.00198 **
Polio_transformed              8.278e-05  1.636e-04   0.506  0.61291
Diphtheria_transformed         2.105e-04  1.372e-04   1.535  0.12497
Status2                       -3.597e+03  1.196e+03  -3.008  0.00267 **
IncomeComposition_transformed  2.370e+05  5.696e+03  41.598  < 2e-16 ***
BMI_transformed                3.613e+01  1.822e+01   1.983  0.04754 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16160 on 1826 degrees of freedom
Multiple R-squared:  0.8121,    Adjusted R-squared:  0.8115
F-statistic:  1315 on 6 and 1826 DF,  p-value: < 2.2e-16
```

Adjusted R-squared for the transfored model is 81.43%, which indicates that the model explains a significant proportion of the variance, even when accounting for the number of predictors, and the p-value of $< 2.2e\text{-}16$ suggests that the model is highly statistically significant overall.

After the transformation, all the distribution predictors satisfy the assumption of constant vriance, uncorrelated error, and linearity.

The QQ-plot of residuals are closer to the 45-degree with less deviation after the transformation, which follows the assumption of normality.

The distribution of residuals is still normalized, which satisfies the assumption of normality. Until now, all the assumptions are followed, we can continue to model reduction using AIC

```
Start:  AIC=35530.84
LifeExpectancy_transformed ~ GDP_transformed + Diphtheria_transformed +
    Polio_transformed + IncomeComposition_transformed + Status +
    BMI_transformed
```

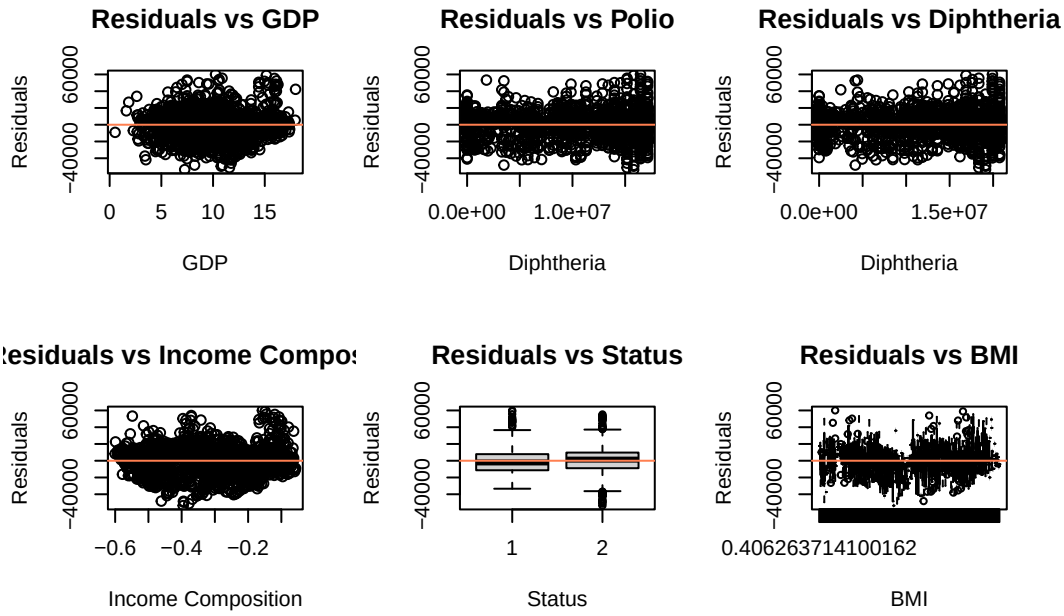

Figure 7: Residuals versus Each Predictor

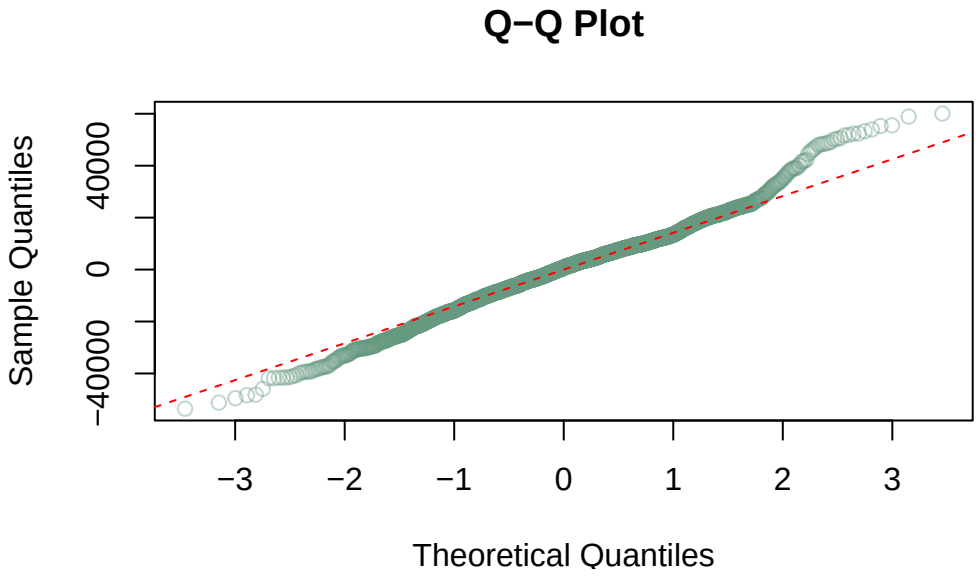

Figure 8: Q-Q Plot of Residuals

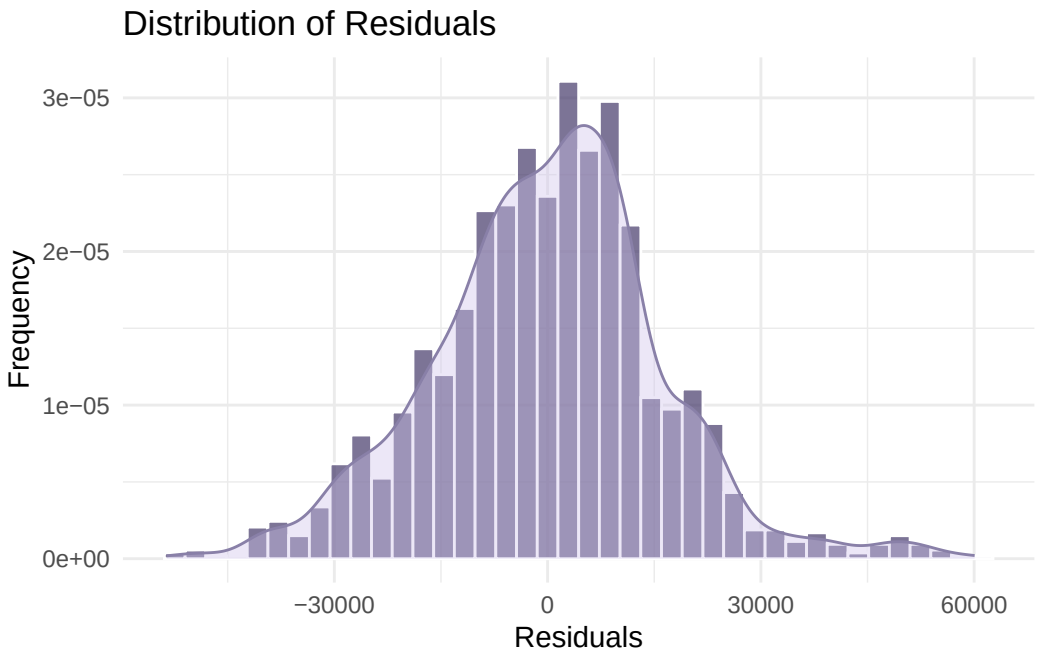

Figure 9: Distribution of Response Value

```
                                Df  Sum of Sq        RSS   AIC
- Polio_transformed              1 6.6838e+07 4.7672e+11 35529
<none>                                       4.7665e+11 35531
- Diphtheria_transformed         1 6.1501e+08 4.7727e+11 35531
- BMI_transformed                1 1.0263e+09 4.7768e+11 35533
- Status                         1 2.3619e+09 4.7902e+11 35538
- GDP_transformed                1 2.5043e+09 4.7916e+11 35538
- IncomeComposition_transformed  1 4.5170e+11 9.2836e+11 36751

Step:  AIC=35529.1
LifeExpectancy_transformed ~ GDP_transformed + Diphtheria_transformed +
    IncomeComposition_transformed + Status + BMI_transformed

                                Df  Sum of Sq        RSS   AIC
<none>                                       4.7672e+11 35529
- BMI_transformed                1 1.0194e+09 4.7774e+11 35531
- Status                         1 2.3493e+09 4.7907e+11 35536
- GDP_transformed                1 2.5177e+09 4.7924e+11 35537
- Diphtheria_transformed         1 3.6102e+09 4.8033e+11 35541
- IncomeComposition_transformed  1 4.6030e+11 9.3702e+11 36766
```

```
Call:
lm(formula = LifeExpectancy_transformed ~ GDP_transformed + Diphtheria_transformed +
    IncomeComposition_transformed + Status + BMI_transformed,
    data = data_clean)

Coefficients:
                  (Intercept)                GDP_transformed
                    1.765e+05                     -5.284e+02
       Diphtheria_transformed  IncomeComposition_transformed
                    2.695e-04                      2.373e+05
                      Status2                BMI_transformed
                   -3.587e+03                      3.601e+01
```

AIC would select a better fit of the model. Based on the backward elimination procedure (See **?@fig-aic**), all AIC of sub-models are larger than the initial model, none of the predictors should be removed from the initial model and all predictors (`GDP`, `Diphtheria`, `Status`, `BMI`, and `IncomeComposition`) contribute significantly to explaining the variation in `LifeExpectancy` supported by this statistical evidence.

The predictor `Polio_transformed` needs to be dropped from the model because its p-value is high (0.646), indicating that it is not statistically significant in explaining the variation in life expectancy. Additionally, removing `Polio_transformed` leads to an improvement as reflected by the AIC. This suggests that `Polio_transformed` does not provide valuable information for the model and may be redundant, possibly due to multicollinearity with other predictors like `Diphtheria_transformed`. By removing `Polio_transformed`, the model becomes simpler and more focused on the predictors that have a stronger impact on life expectancy.

```
              GDP_transformed        Diphtheria_transformed
                     2.107062                      1.604884
IncomeComposition_transformed                        Status
                     4.120550                      1.550791
              BMI_transformed
                     1.675942
```

In the result of VIF((**git-vif?**)), all the VIF values are below 5, which suggests that there is no significant multicollinearity among the predictors in the reduced model. This means that the variables are relatively independent of each other

After the transformation, all the distribution predictors satisfy the assumption of constant vriance, uncorrelated error, and linearity. The QQ-plot of residuals are closer to the 45-degree with less deviation after the transformation, which follows the assumption of normality.

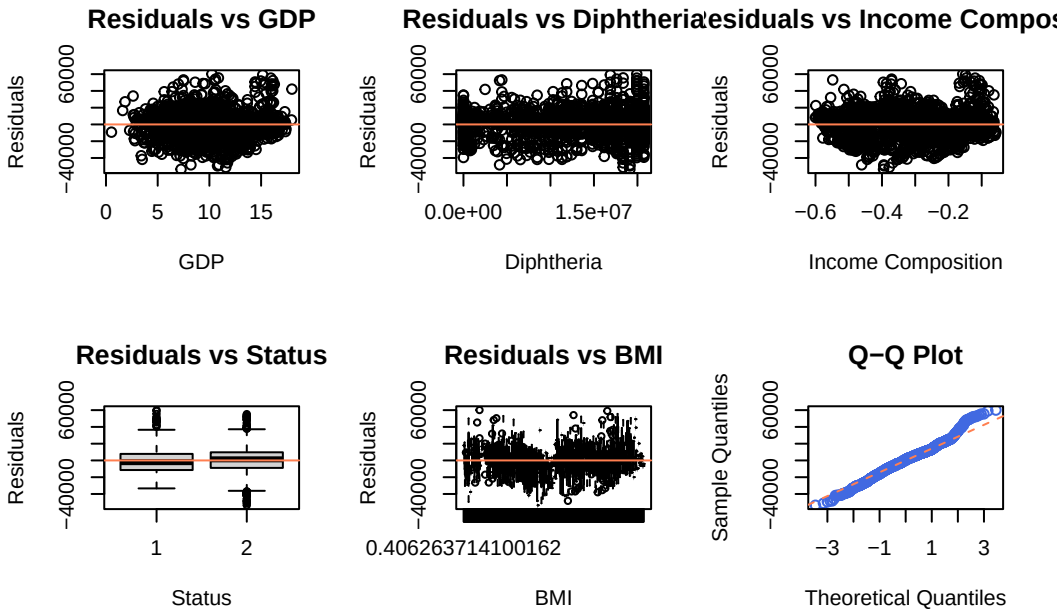

Figure 10: Residuals versus Each Predictor

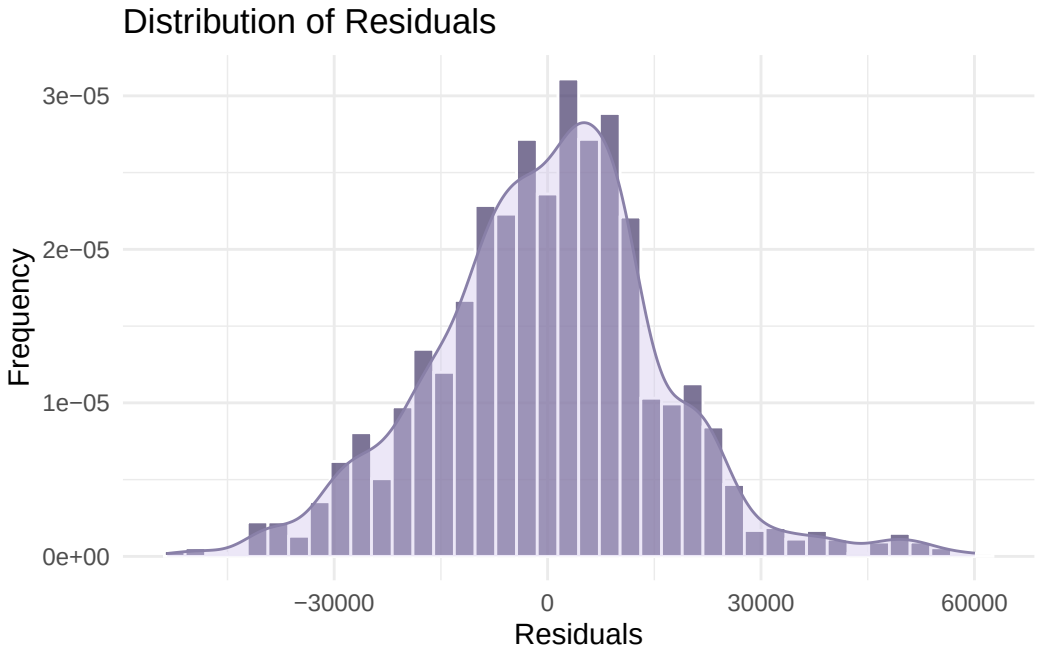

Figure 11: Distribution of Response Value

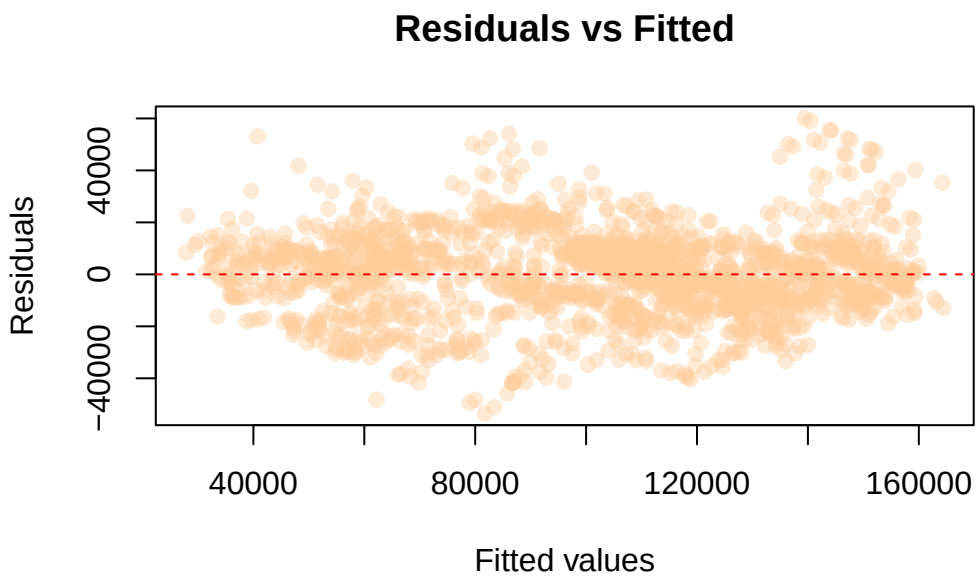

Figure 12: Residuals vs Fitted after transformation and fitting transformed model

The distribution of residuals is normalized, which satisfies the assumption of normality

the residuals are evenly scattered around 0, with no discernible pattern. This suggests that the linearity assumption is met. So far, all the assumptions are checked and satisfied.

```
Analysis of Variance Table

Model 1: LifeExpectancy_transformed ~ GDP_transformed + Diphtheria_transformed +
    IncomeComposition_transformed + Status + BMI_transformed
Model 2: LifeExpectancy_transformed ~ GDP_transformed + Polio_transformed +
    Diphtheria_transformed + Status + IncomeComposition_transformed +
    BMI_transformed
  Res.Df        RSS Df Sum of Sq     F Pr(>F)
1   1827 4.7672e+11
2   1826 4.7665e+11  1  66837721 0.256 0.6129
```

Now we apply ANOVA to compare the reduced model, which excludes `Polio_transformed`, with the transformed model that includes all the predictors. The F-statistic is 0.2105, and the p-value is 0.6464, which is much higher than the significance value 0.05 indicating that it fails to reject the null hypothesis In other words, `Polio_transformed` is not significantly contributing to the explanation of `LifeExpectancy_transformed`. Therefore, there is evidence that the `reduced_fit` fits better than `transformed_fit.`

3 Results

Table 1: Predicted average Life Extectancy for developed and developing countries based on socioeconomic factors

Average Life Expectancy in Developed Countries	Average Life Expectancy in Developing Countries
67.04	68.87

```
ggplot(data_clean, aes(x = TotalExpenditure, y = LifeExpectancy, color = Status)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Year", y = "Life Expectancy")
```

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

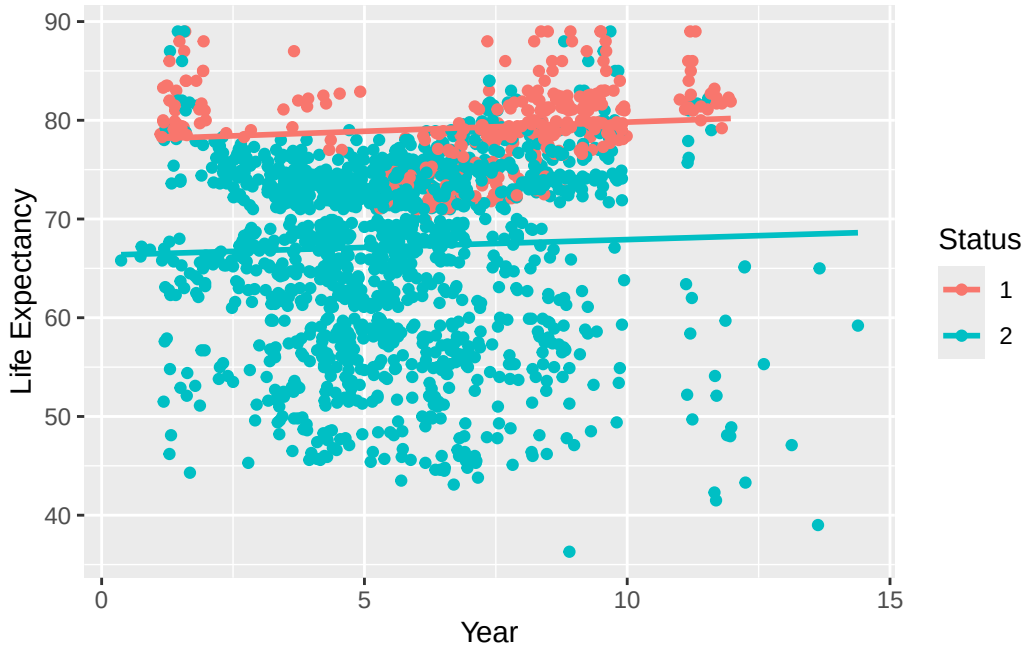

Table 2: Summary of key model estimates for Developed and Developing counties, including coefficients for predictors like Diphtheria, GDP, TotalExpenditure, and BMI, with standard errors for each estimate. Model performance statistics, such as sample size, R^2, and adjusted R^2, are also displayed.

	Model for Developed	Model for Developing
Diphtheria Immunization	0.024	0.078
	(0.010)	(0.006)
GDP	0.000	0.000
	(0.000)	(0.000)
Income Composition	54.731	17.240
	(3.076)	(0.876)
BMI	−0.009	0.142
	(0.008)	(0.009)
Num.Obs.	420	1881
R2	0.513	0.550
R2 Adj.	0.509	0.549
AIC	2070.3	12 189.2
BIC	2094.5	12 222.4
Log.Lik.	−1029.140	−6088.586
RMSE	2.80	6.16

Table 3: Summary of the Life Expectancy model, which includes key variables such as GDP, Diphtheria, income composition The table presents the model coefficients along with their standard errors.

Term	Estimate	Std. Error	t Value	Pr(>|t|)	Term_Type
Intercept	51.393	1.014	50.669	0	Linear
Status	-2.935	0.385	-7.617	0	Linear
GDP	0.000	0.000	6.669	0	Linear
BMI	0.114	0.007	15.579	0	Linear
Diphtheria	0.072	0.006	12.547	0	Linear
IncomeComposition	19.711	0.796	24.752	0	Linear

The summary of the Life Expectancy model, which includes key variables such as GDP, Diphtheria, income composition, highlights the contribution of each predictor to the multiple linear regression model. The `Intercept` represents the predicted baseline life expectancy when all other predictors are zero. where status GDP performs a nuanced contribution to the model as shown in the table, this is because GDP is a measurement of the value of goods and services bought and sold in markets, so it

4 Discussion

4.1 First discussion point

In the response vs. 5 predictors plots (Figure 1), Status categorizes countries into developed or developing status, which distincts countries by their economic dependence and medical development level. GDP (in USD) is the statistic that measures the economic welfare of a country. Income composition of resources is the HDI measurement on the standard of living calculated by GNI per capita, which is commonly used to reflect average income. Polio and Diphtheria represent the immunization coverage among 1-year-olds children in percentage and they manifest the extensiveness and accessibility for public health, where widespread infectious diseases can potentially lower life expectancy as these two diseases are fatal to children. The above five predictors reveal the proportional relationships between national economic standings and life expectancy and analyze whether higher economic standings countries have a larger life expectancy.

4.2 Second discussion point - Preliminary Results (212 words)

From the Residuals vs Fitted graph(Figure 12) and the plots of Residuals vs each Predictor (**?@fig-residuals-plots**), the residuals appear randomly scattered around the zero line, which satisfies linearity There is no discernible pattern in the residuals, suggesting the model captures the linear relationship between the predictors and response variable well. The residuals also do not display any systematic pattern, implying that the errors are uncorrelated.

In QQ Plot(Figure 5), the residuals generally follow the 45-degree reference line, which suggests that the assumption of normality is mostly satisfied. However, the points at both ends deviate from the line, indicating potential outliers. The presence of outliers requires further transformation of both response and predictors values

The Scale-Location graph(**?@fig-scale-location**) shows that while the points are not fully evenly dispersed, they do not exhibit a clear pattern or trend, indicating that the assumption of Constant Variance may not be entirely satisfied. The presence of a non-horizontal line suggests variability in the spread of residuals, which our team should address in the next steps to improve model accuracy.

The Residuals vs. Leverage plot(**?@fig-leverage**) shows some points clustering around zero, while others are more dispersed, with no clear pattern evident. This indicates a need for our team to improve the distribution of points in future analyses.

4.3 Weaknesses and next steps

5 Conclusion

Appendix

A Additional data details

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.