**Exploring Global Life Expectancy Patterns and Predictors: A Comprehensive Analysis of the WHO Dataset**

**Group Name**: A Plus

**Name**: Yi Yang, Weiyi Gong, Xiaolong Wang, Zeming Ren

**Student ID:** yy3735, wg733, xw2893, zr2173

**Contributors**:

YY & WG are responsible for conception, design, cleaning, and analysis of the data. YY are responsible for the codebook. YY, WG, XW, and ZR drafted the article. WG conducted an introduction of the article. ZR conducted the descriptive and method section of the article. XW conducted the result section of the article. YY conducted the conclusion section of the article. All authors read and approved the final draft of the manuscript.

1. **Introduction**

*1.1 Background*

Life expectancy, which is defined as the average number of years a newborn is expected to live, is a crucial indicator of a population's overall health, well-being, and quality of life (World Health Organization, 2021).Specifically, we determined how body mass index, demographic variables, and physical variables were associated with the length of life expectancy comparing developed and developing countries.

*1.2 Models*

1.2.1.1 Model 1

Prior research has shown a strong correlation between body mass index (BMI) and life expectancy, with both high (obesity) and low (underweight) BMI levels associated with increased mortality rates (Prospective Studies Collaboration, 2009). In Model 1, we hypothesized that body mass index, adult mortality rate, infant deaths rate, under-five death rate, HIV/AIDS rate, income composition of resources, and schooling rate were positively associated with the length of life expectancy.

1.2.2 Model 2a

In developed countries, the prevalence of obesity has increased considerably over the past few decades (Ng et al., 2014), resulting in increased risks of non-communicable diseases such as cardiovascular diseases and certain types of cancer, which can have a negative effect on life expectancy (Guh et al., 2009). In developing countries, on the other hand, undernutrition and low BMI continue to be significant public health concerns (Black et al., 2013), contributing to higher mortality rates, especially among children (Olofin et al., 2013).

In Model 2a, we included all the predictors in Model 1 with additional interaction terms of BMI and variable developing with life expectancy.

1.2.3 Model 2b

In Model 2b, we included all the predictors in Model 1 with additional interaction terms of BMI and expenditure on health with life expectancy.

1.2.4 Model 3

Access to healthcare, socioeconomic conditions, and public health policies can impact the relationship between BMI and life expectancy (Swinburn et al., 2011).

In Model 3, we hypothesized that all the variables selected were positively associated with life expectancy.

1. **Methods**

*2.1 Descriptive Statistics*

The data used in this analysis project is a life expectancy project, which has 22 variables and 2938 observations. In the data cleaning process, we removed several useless variables and missing rows. The final data we used consisted of 20 variables and 1649 observations. The 20 variables are life expectancy, adult mortality between 15 and 60 years per 1000, infant death number, alcohol consumption in liter, Expenditure on health as a percentage of GDP, Hepatitis B immune rate, measles cases per 1000, BMI, under 5 deaths per 1000, polio immune coverage among 1 year old, general government expenditure on health as percentage of total government expenditure, Diphtheria tetanus toxoid, and pertussis (DTP3) immunization coverage among 1-year-olds, Deaths per 1000 live births HIV/AIDS in 0-4 years old, GDP, population, Prevalence of thinness among children and adolescents for Age 10 to 19, Prevalence of thinness among children and adolescents for Age 5 to 9, Human Development Index (from 0 to 1), years of schooling, and Recode Developed or Developing status.

Due to the fact that the variables selected are all continuous variables, the summary of variables is shown below as mean and standard deviation in Table 1. It could help us learn more about the distribution of variables and check for outliers. From the summary table, the mean life expectancy is 69.3, the median is 71.7, and it is distributed from 44 to 89. The mean BMI is 38.13, the median is 43.7, and it is distributed from 2 to 77.1.

*2.2 Bivariate Analysis*

The bivariate analysis table shows the correlation between the main outcome (life expectancy) and main exposure (BMI) with all variables. Observe variables with high correlation values and consider adding interaction terms to improve the accuracy of the models.

*2.3 Multivariable Regression*

In this project, several models were performed to predict life expectancy, and the models' mean squared error (MSE) was compared to select the best model. The MSE tells us how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them(Statisticshowto (n.d.), 2023). At first, researchers fitted a full model and checked the p-value of each variable. Model 1 uses all significant variables in the full model; it has seven variables, which contain BMI, adult mortality, infant deaths, under-five years mortality, HIV/AIDS rate, human development index, and school years. From the bivariate analysis, we found that variable developing has a high correlation with BMI and life expectancy. Model 2a is based on the first model; the 8 variables consist of all variables in the first model plus an interaction term with BMI and developing. Model 2b consists of all the variables in Model 1 plus the interaction term between BMI and expenditure. For Model 3, we first fitted a linear regression model on all the variables and found all the significant predictors with p-values less than 0.05, then we used selection methods including AIC, BIC, Lasso, adaptive Lasso, and ridge. MSE was used as an assessment measurement for model comparison.

1. **Results**

Table 1: Univariate Analysis

| Characteristic | Mean±SD(range) |
| --- | --- |
| Life.expectancy | 69.30±8.80 (68.88, 69.73) |
| Adult.Mortality | 168.22±125.31 (162.16, 174.77) |
| infant.deaths | 32.55±120.85 (26.72, 38.39) |
| Alcohol | 4.53±4.03 (4.34, 4.73) |
| percentage.expenditure | 698.97±1759.23 (614.00, 783.95) |
| Hepatitis.B | 79.22±25.60 (77.98, 80.45) |
| Measles | 2224.49±10085.80 (1737.34, 2711.65) |
| BMI | 38.13±19.75 (37.17, 39.08) |
| under.five.deaths | 44.22±162.90 (36.35, 52.09) |
| Polio | 83.56±22.45 (82.48, 84.65) |
| Total.expenditure | 5.96±2.30 (5.84, 6.07) |
| Diphtheria | 84.16±21.58 (83.11, 85.20) |
| HIV.AIDS | 1.98±6.03 (1.69, 2.28) |
| GDP | 5566.03±11475.90 (5011.73, 6120.33) |
| thinness..1.19.years | 4.85±4.60 (4.63, 5.07) |
| thinness.5.9.years | 4.91±4.65 (4.68, 5.13) |
| Income.composition.of.resources | 0.63±0.18 (0.62, 0.64) |
| Schooling | 12.12±2.80 (11.98, 12.25) |
| Developing | 0.85±0.35 (0.84, 0.87) |

Table 2: Bivariate Analysis

|  | Life.expectancy |  | BMI | |
| --- | --- | --- | --- | --- |
|  | cor | p | cor | p |
| Life.expectancy | - | - | 0.54 | 0.00 |
| BM BMI | 0.54 | 0.00 | - | - |
| Adult.Mortality | -0.70 | 0.00 | -0.35 | 0.00 |
| infant.deaths | -0.17 | 0.00 | -0.23 | 0.00 |
| Alcohol | 0.40 | 0.00 | 0.35 | 0.00 |
| percentage.expenditure | 0.40 | 0.00 | 0.24 | 0.00 |
| Hepatitis.B | 0.20 | 0.00 | 0.14 | 0.00 |
| Measles | -0.07 | 0.01 | -0.15 | 0.00 |
| under.five.deaths | -0.19 | 0.00 | -0.24 | 0.00 |
| Polio | 0.33 | 0.00 | 0.19 | 0.00 |
| Total.expenditure | 0.17 | 0.00 | 0.19 | 0.00 |
| Diphtheria | 0.34 | 0.00 | 0.18 | 0.00 |
| HIV.AIDS | -0.59 | 0.00 | -0.21 | 0.00 |
| GDP | 0.44 | 0.00 | 0.27 | 0.00 |
| Population | -0.02 | 0.37 | -0.08 | 0.00 |
| thinness..1.19.years | -0.46 | 0.00 | -0.55 | 0.00 |
| thinness.5.9.years | -0.46 | 0.00 | -0.55 | 0.00 |
| Income.composition.of.resources | 0.72 | 0.00 | 0.51 | 0.00 |
| Schooling | 0.73 | 0.00 | 0.56 | 0.00 |
| Developing | -0.44 | 0.00 | -0.30 | 0.00 |

Figure 1: Original and transformed distribution of life expectancy

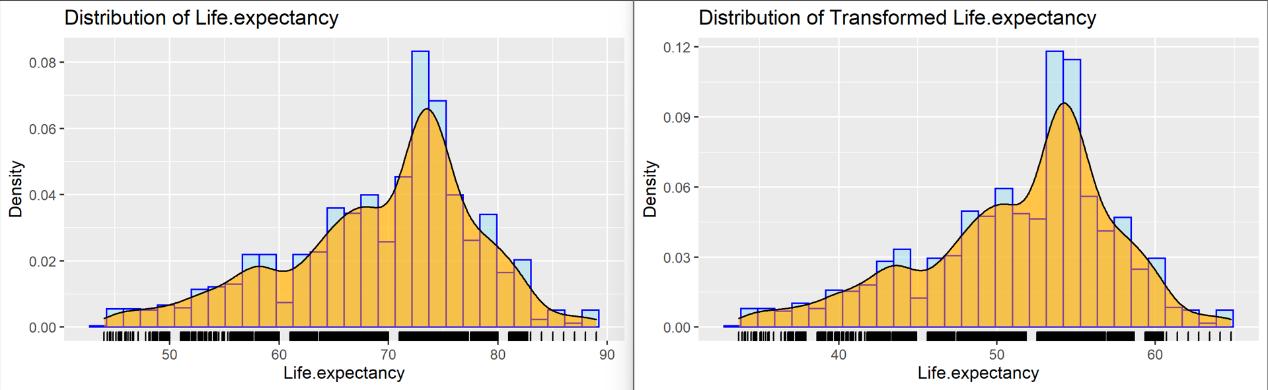


Table 3: Results of several constructed models

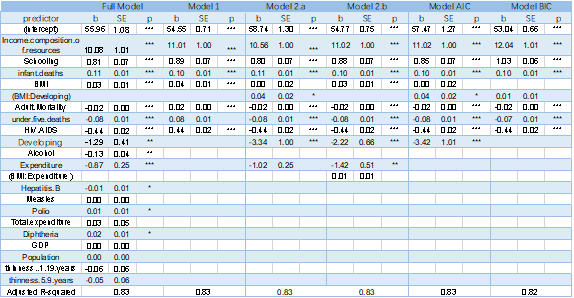
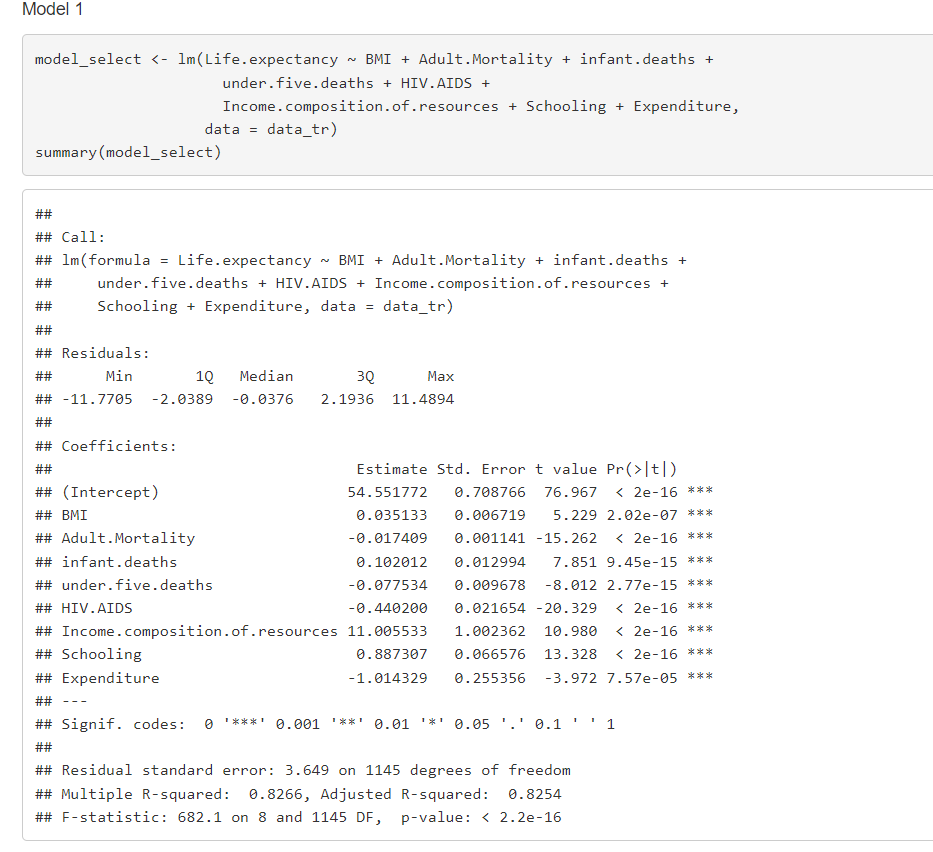
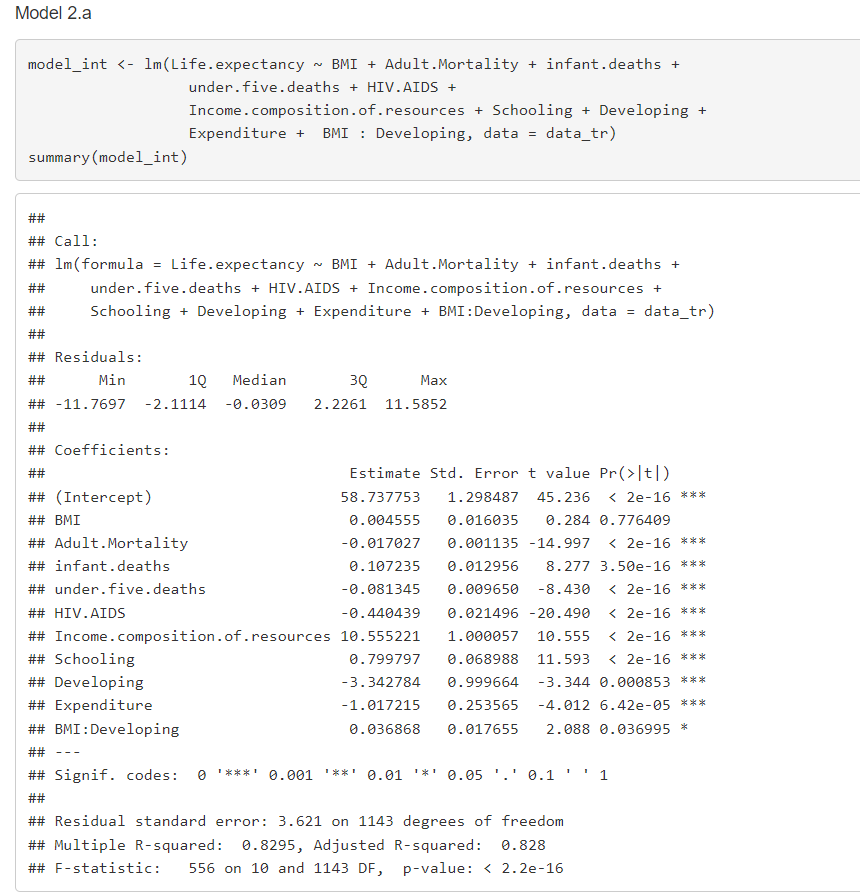


Figure 1 presents the distribution of life expectancy, and all samples are deemed reliable and accurate. The dependent variable is approximately normally distributed, with a mean of 69.30 and a variance of 8.80. Table I provides the outcomes of the univariate analysis of all variables, and a comprehensive summary is presented in the same table. Moreover, Table 2 provides the bivariate analysis results. Based on the definition of correlation and prior research design(Falvo & Earhart ,2009), we can conclude that all variables, except population, are correlated with BMI and life expectancy. And based on the format of the writing sample(Palamar, Kiang & Halkitis, 2012), we draw Table 3 to present the summary of results.

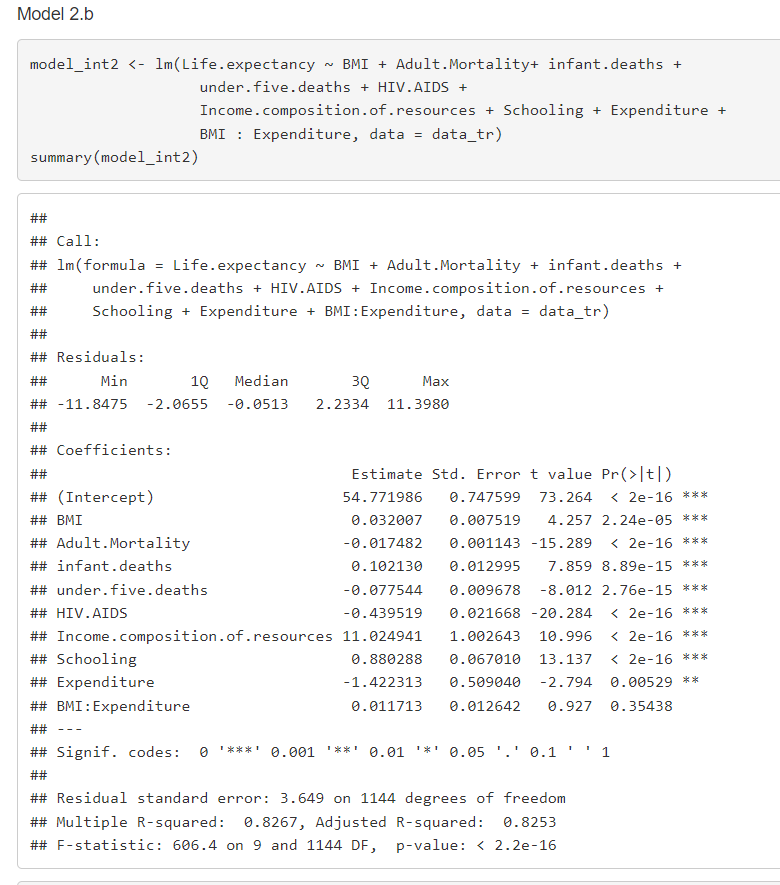
Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no errors, the MSE equals zero.



A concise Model 1 was developed using several variables of interest based on the full predict model. Model 1 had only six predictor variables, yet it had a similar r-square value (0.8254, p < 0.001) to the full model, as shown in Table 3. By Occam's razor principle, Model 1 was selected as the baseline for this study.

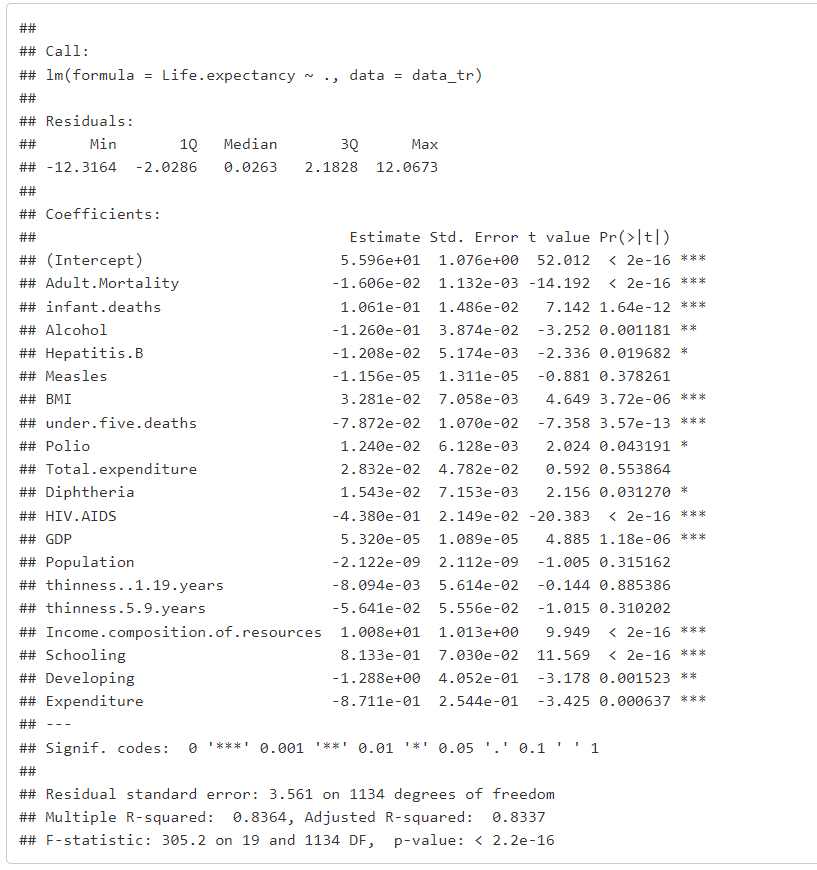


Several attempts were made to enhance the existing Model 1. As a result, the Model 2’s r-squared improved to (0.828, p < 0.001) , and BMI was no longer significant, while the interaction term became significant. This means that the effect of BMI on life expectancy depends on the variable developing. In other words, people with the same BMI may have a 0.04 higher life expectancy in a developed country than if they were in a developing country. This led us to conclude that Model 2a was superior to Model 1.

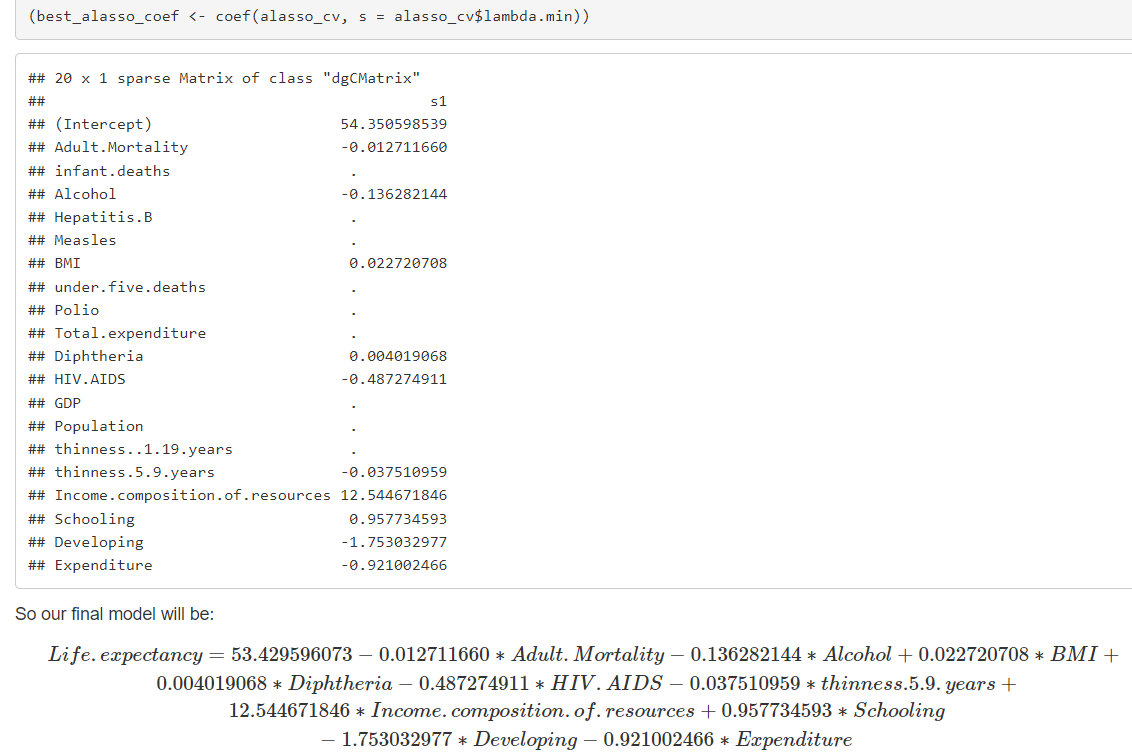


Building upon Model 2a, we examine the effect of BMI on life expectancy depending on the level of expenditure in Model 2b. However, the new model (adjusted R squared = 0.825, p < 0.001) did not show a significant effect of the interaction between BMI and health expenditure.

For the AIC and BIC methods, our findings indicate that the model generated by the AIC method (adjusted R squared = 0.825, p < 0.001) was consistent with Model 2a. On the other hand, the model generated by the BIC method (adjusted R squared = 0.819, p < 0.001) yielded poorer results than Model 2a.



The adjusted r-squared for our model 3 is 0.8337. Adult mortality, Adult.Mortality, infant.deaths, BMI, under.five.deaths, HIV/AIDS, GDP, Income.composition.of.resources, Schooling, Expenditure, Alcohol, Hepatitis.B, Polio, Diphtheria, Developing all have p-values < 0.05, thus concluding that they are all significant to life expectancy. Results of the omnibus test: F(19, 1134) = 305.2; the p-value is less than 0.001.



1. **Conclusions**

Our group has selected a final model using the AIC, BIC, Lasso, Ridge, and Adjust Lasso models. AIC and BIC are useful for selecting the best model from a set of candidate models, while Lasso and Ridge regressions are useful for variable selection and can help prevent overfitting in situations where there are a large number of predictor variables.

After our observations, we confirmed that body mass index (BMI), demographic variables, and physical variables have positive associations with the length of life expectancy comparing developed and developing countries; however, expenditure on health as a percentage of gross domestic product per capita and life expectancy is insignificant.

However, even with this model, we must acknowledge its inherent limitations, including parameter selection, handling outliers and influence, and lack of stability. Therefore, our further research is to ensure algorithmic fairness in the data collection. According to Suresh, "fairness cannot be achieved perfectly and instead is a balance between competing values that must be continuously evaluated and adjusted." (Suresh & Guttag, 2019, p. 32) and also we could try to use a more efficient mode, and the data for each country could be highly reputable, so we could do longitudinal analysis to estimate the predictor variable using the Gee model.

**Reference**

**Introduction**

1. Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., de Onis, M., Ezzati, M., Grantham-McGregor, S., Katz, J., & Martorell, R. (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. The Lancet, 382(9890), 427-451.<https://doi.org/10.1016/S0140-6736(13)60937-X>
2. Guh, D. P., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C. L., & Anis, A. H. (2009). The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis. BMC Public Health, 9, 88.<https://doi.org/10.1186/1471-2458-9-88>
3. Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., Mullany, E. C., Biryukov, S., Abbafati, C., Abera, S. F., & Abraham, J. P. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: A systematic analysis for the Global Burden of Disease Study 2013. The Lancet, 384(9945), 766-781.<https://doi.org/10.1016/S0140-6736(14)60460-8>
4. Olofin, I. O., McDonald, C. M., Ezzati, M., Flaxman, S., Black, R. E., & Fawzi, W. W. (2013). Associations of suboptimal growth with all-cause and cause-specific mortality in children under five years: A pooled analysis of ten prospective studies. PLoS One, 8(5), e64636.<https://doi.org/10.1371/journal.pone.0064636>
5. Prospective Studies Collaboration. (2009). Body-mass index and cause-specific mortality in 900 000 adults: Collaborative analyses of 57 prospective studies. The Lancet, 373(9669), 1083-1096.<https://doi.org/10.1016/S0140-6736(09)60318-4>
6. Swinburn, B. A., Sacks, G., Hall, K. D., McPherson, K., Finegood, D. T., Moodie, M. L., Gortmaker, S. L., & the Obesity Policy Action (OPA) network. (2011). The global obesity pandemic: Shaped by global drivers and local environments. The Lancet, 378(9793), 804-814.<https://doi.org/10.1016/S0140-6736(11)60813-1>
7. World Health Organization. (2021). Life expectancy.<https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/life-expectancy>

**Methods**

1. Statisticshowto. (n.d.). Mean Squared Error (MSE) in Statistics. Retrieved from <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/#MSE>

**Results**

1. Falvo, M. J., & Earhart, G. M. (2009). Six-minute walk distance in persons with Parkinson disease: a hierarchical regression model. Archives of physical medicine and rehabilitation, 90(6), 1004–1008. <https://doi.org/10.1016/j.apmr.2008.12.018>
2. Palamar, J. J., Kiang, M. V., & Halkitis, P. N. (2012). Predictors of stigmatization towards use of various illicit drugs among emerging adults. Journal of psychoactive drugs, 44(3), 243–251. https://doi.org/10.1080/02791072.2012.703510

**Conclusion**

1. Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 30-38). ACM.
2. **Supplementary Documentation**

Regression 1 - Final Project Life Exp - Team A Plus

Yi Yang, Weiyi Gong, Xiaolong Wang, Zeming Ren

2023-04-15

## Data Preparation and cleaning

## 1.Data Cleaning and Descriptive

my\_data <- read.csv("Life Expectancy Data.csv")  
my\_data1 <- my\_data %>%   
 na.omit() %>%  
 mutate(Developing = as.integer(Status == "Developing")) %>%   
 *# Change status to numeric*  
 mutate(Expenditure = as.integer(percentage.expenditure < 65))  
my\_data1 <- my\_data1[, -c(1, 2, 3, 8)] *# remove country, year, status*  
pander(summary(my\_data1), caption = 'Descriptive Statistics of The Data')

*Descriptive Statistics of The Data (continued below)*

| Life.expectancy | Adult.Mortality | infant.deaths | Alcohol |
| --- | --- | --- | --- |
| Min. :44.0 | Min. : 1.0 | Min. : 0.00 | Min. : 0.010 |
| 1st Qu.:64.4 | 1st Qu.: 77.0 | 1st Qu.: 1.00 | 1st Qu.: 0.810 |
| Median :71.7 | Median :148.0 | Median : 3.00 | Median : 3.790 |
| Mean :69.3 | Mean :168.2 | Mean : 32.55 | Mean : 4.533 |
| 3rd Qu.:75.0 | 3rd Qu.:227.0 | 3rd Qu.: 22.00 | 3rd Qu.: 7.340 |
| Max. :89.0 | Max. :723.0 | Max. :1600.00 | Max. :17.870 |

*Table continues below*

| Hepatitis.B | Measles | BMI | under.five.deaths |
| --- | --- | --- | --- |
| Min. : 2.00 | Min. : 0 | Min. : 2.00 | Min. : 0.00 |
| 1st Qu.:74.00 | 1st Qu.: 0 | 1st Qu.:19.50 | 1st Qu.: 1.00 |
| Median :89.00 | Median : 15 | Median :43.70 | Median : 4.00 |
| Mean :79.22 | Mean : 2224 | Mean :38.13 | Mean : 44.22 |
| 3rd Qu.:96.00 | 3rd Qu.: 373 | 3rd Qu.:55.80 | 3rd Qu.: 29.00 |
| Max. :99.00 | Max. :131441 | Max. :77.10 | Max. :2100.00 |

*Table continues below*

| Polio | Total.expenditure | Diphtheria | HIV.AIDS |
| --- | --- | --- | --- |
| Min. : 3.00 | Min. : 0.740 | Min. : 2.00 | Min. : 0.100 |
| 1st Qu.:81.00 | 1st Qu.: 4.410 | 1st Qu.:82.00 | 1st Qu.: 0.100 |
| Median :93.00 | Median : 5.840 | Median :92.00 | Median : 0.100 |
| Mean :83.56 | Mean : 5.956 | Mean :84.16 | Mean : 1.984 |
| 3rd Qu.:97.00 | 3rd Qu.: 7.470 | 3rd Qu.:97.00 | 3rd Qu.: 0.700 |
| Max. :99.00 | Max. :14.390 | Max. :99.00 | Max. :50.600 |

*Table continues below*

| GDP | Population | thinness..1.19.years |
| --- | --- | --- |
| Min. : 1.68 | Min. :3.400e+01 | Min. : 0.100 |
| 1st Qu.: 462.15 | 1st Qu.:1.919e+05 | 1st Qu.: 1.600 |
| Median : 1592.57 | Median :1.420e+06 | Median : 3.000 |
| Mean : 5566.03 | Mean :1.465e+07 | Mean : 4.851 |
| 3rd Qu.: 4718.51 | 3rd Qu.:7.659e+06 | 3rd Qu.: 7.100 |
| Max. :119172.74 | Max. :1.294e+09 | Max. :27.200 |

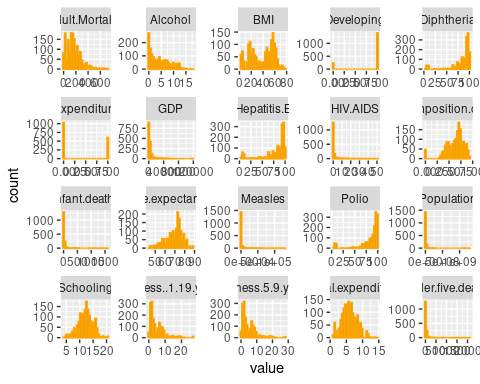
*Table continues below*

| thinness.5.9.years | Income.composition.of.resources | Schooling |
| --- | --- | --- |
| Min. : 0.100 | Min. :0.0000 | Min. : 4.20 |
| 1st Qu.: 1.700 | 1st Qu.:0.5090 | 1st Qu.:10.30 |
| Median : 3.200 | Median :0.6730 | Median :12.30 |
| Mean : 4.908 | Mean :0.6316 | Mean :12.12 |
| 3rd Qu.: 7.100 | 3rd Qu.:0.7510 | 3rd Qu.:14.00 |
| Max. :28.200 | Max. :0.9360 | Max. :20.70 |

| Developing | Expenditure |
| --- | --- |
| Min. :0.0000 | Min. :0.0000 |
| 1st Qu.:1.0000 | 1st Qu.:0.0000 |
| Median :1.0000 | Median :0.0000 |
| Mean :0.8532 | Mean :0.3699 |
| 3rd Qu.:1.0000 | 3rd Qu.:1.0000 |
| Max. :1.0000 | Max. :1.0000 |

my\_data1 %>%  
 keep(is.numeric) %>%   
 gather() %>%   
 ggplot(aes(value)) +  
 facet\_wrap(~ key, scales = "free") +  
 geom\_histogram(col = 'orange')

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



pander(skewness(my\_data1), caption = 'Skewness of numeric data')

*Table continues below*

| Life.expectancy | Adult.Mortality | infant.deaths | Alcohol | Hepatitis.B |
| --- | --- | --- | --- | --- |
| -0.6282 | 1.275 | 8.47 | 0.6619 | -1.792 |

*Table continues below*

| Measles | BMI | under.five.deaths | Polio | Total.expenditure |
| --- | --- | --- | --- | --- |
| 7.951 | -0.2334 | 8.333 | -2.358 | 0.2132 |

*Table continues below*

| Diphtheria | HIV.AIDS | GDP | Population | thinness..1.19.years |
| --- | --- | --- | --- | --- |
| -2.485 | 4.97 | 4.513 | 14.17 | 1.819 |

*Table continues below*

| thinness.5.9.years | Income.composition.of.resources | Schooling | Developing |
| --- | --- | --- | --- |
| 1.865 | -1.154 | -0.128 | -1.997 |

| Expenditure |
| --- |
| 0.5389 |

pander(head(my\_data1), caption = 'First six rows of data')

*First six rows of data (continued below)*

| Life.expectancy | Adult.Mortality | infant.deaths | Alcohol | Hepatitis.B |
| --- | --- | --- | --- | --- |
| 65 | 263 | 62 | 0.01 | 65 |
| 59.9 | 271 | 64 | 0.01 | 62 |
| 59.9 | 268 | 66 | 0.01 | 64 |
| 59.5 | 272 | 69 | 0.01 | 67 |
| 59.2 | 275 | 71 | 0.01 | 68 |
| 58.8 | 279 | 74 | 0.01 | 66 |

*Table continues below*

| Measles | BMI | under.five.deaths | Polio | Total.expenditure | Diphtheria |
| --- | --- | --- | --- | --- | --- |
| 1154 | 19.1 | 83 | 6 | 8.16 | 65 |
| 492 | 18.6 | 86 | 58 | 8.18 | 62 |
| 430 | 18.1 | 89 | 62 | 8.13 | 64 |
| 2787 | 17.6 | 93 | 67 | 8.52 | 67 |
| 3013 | 17.2 | 97 | 68 | 7.87 | 68 |
| 1989 | 16.7 | 102 | 66 | 9.2 | 66 |

*Table continues below*

| HIV.AIDS | GDP | Population | thinness..1.19.years | thinness.5.9.years |
| --- | --- | --- | --- | --- |
| 0.1 | 584.3 | 33736494 | 17.2 | 17.3 |
| 0.1 | 612.7 | 327582 | 17.5 | 17.5 |
| 0.1 | 631.7 | 31731688 | 17.7 | 17.7 |
| 0.1 | 670 | 3696958 | 17.9 | 18 |
| 0.1 | 63.54 | 2978599 | 18.2 | 18.2 |
| 0.1 | 553.3 | 2883167 | 18.4 | 18.4 |

| Income.composition.of.resources | Schooling | Developing | Expenditure |
| --- | --- | --- | --- |
| 0.479 | 10.1 | 1 | 0 |
| 0.476 | 10 | 1 | 0 |
| 0.47 | 9.9 | 1 | 0 |
| 0.463 | 9.8 | 1 | 0 |
| 0.454 | 9.5 | 1 | 1 |
| 0.448 | 9.2 | 1 | 0 |

pander(summary(my\_data1), caption = 'First six rows of data')

*First six rows of data (continued below)*

| Life.expectancy | Adult.Mortality | infant.deaths | Alcohol |
| --- | --- | --- | --- |
| Min. :44.0 | Min. : 1.0 | Min. : 0.00 | Min. : 0.010 |
| 1st Qu.:64.4 | 1st Qu.: 77.0 | 1st Qu.: 1.00 | 1st Qu.: 0.810 |
| Median :71.7 | Median :148.0 | Median : 3.00 | Median : 3.790 |
| Mean :69.3 | Mean :168.2 | Mean : 32.55 | Mean : 4.533 |
| 3rd Qu.:75.0 | 3rd Qu.:227.0 | 3rd Qu.: 22.00 | 3rd Qu.: 7.340 |
| Max. :89.0 | Max. :723.0 | Max. :1600.00 | Max. :17.870 |

*Table continues below*

| Hepatitis.B | Measles | BMI | under.five.deaths |
| --- | --- | --- | --- |
| Min. : 2.00 | Min. : 0 | Min. : 2.00 | Min. : 0.00 |
| 1st Qu.:74.00 | 1st Qu.: 0 | 1st Qu.:19.50 | 1st Qu.: 1.00 |
| Median :89.00 | Median : 15 | Median :43.70 | Median : 4.00 |
| Mean :79.22 | Mean : 2224 | Mean :38.13 | Mean : 44.22 |
| 3rd Qu.:96.00 | 3rd Qu.: 373 | 3rd Qu.:55.80 | 3rd Qu.: 29.00 |
| Max. :99.00 | Max. :131441 | Max. :77.10 | Max. :2100.00 |

*Table continues below*

| Polio | Total.expenditure | Diphtheria | HIV.AIDS |
| --- | --- | --- | --- |
| Min. : 3.00 | Min. : 0.740 | Min. : 2.00 | Min. : 0.100 |
| 1st Qu.:81.00 | 1st Qu.: 4.410 | 1st Qu.:82.00 | 1st Qu.: 0.100 |
| Median :93.00 | Median : 5.840 | Median :92.00 | Median : 0.100 |
| Mean :83.56 | Mean : 5.956 | Mean :84.16 | Mean : 1.984 |
| 3rd Qu.:97.00 | 3rd Qu.: 7.470 | 3rd Qu.:97.00 | 3rd Qu.: 0.700 |
| Max. :99.00 | Max. :14.390 | Max. :99.00 | Max. :50.600 |

*Table continues below*

| GDP | Population | thinness..1.19.years |
| --- | --- | --- |
| Min. : 1.68 | Min. :3.400e+01 | Min. : 0.100 |
| 1st Qu.: 462.15 | 1st Qu.:1.919e+05 | 1st Qu.: 1.600 |
| Median : 1592.57 | Median :1.420e+06 | Median : 3.000 |
| Mean : 5566.03 | Mean :1.465e+07 | Mean : 4.851 |
| 3rd Qu.: 4718.51 | 3rd Qu.:7.659e+06 | 3rd Qu.: 7.100 |
| Max. :119172.74 | Max. :1.294e+09 | Max. :27.200 |

*Table continues below*

| thinness.5.9.years | Income.composition.of.resources | Schooling |
| --- | --- | --- |
| Min. : 0.100 | Min. :0.0000 | Min. : 4.20 |
| 1st Qu.: 1.700 | 1st Qu.:0.5090 | 1st Qu.:10.30 |
| Median : 3.200 | Median :0.6730 | Median :12.30 |
| Mean : 4.908 | Mean :0.6316 | Mean :12.12 |
| 3rd Qu.: 7.100 | 3rd Qu.:0.7510 | 3rd Qu.:14.00 |
| Max. :28.200 | Max. :0.9360 | Max. :20.70 |

| Developing | Expenditure |
| --- | --- |
| Min. :0.0000 | Min. :0.0000 |
| 1st Qu.:1.0000 | 1st Qu.:0.0000 |
| Median :1.0000 | Median :0.0000 |
| Mean :0.8532 | Mean :0.3699 |
| 3rd Qu.:1.0000 | 3rd Qu.:1.0000 |
| Max. :1.0000 | Max. :1.0000 |

## 2.Define Training and test dataset

set.seed(0)  
tr\_size <- nrow(my\_data1) \* 0.7 *# training sample size*  
tr\_ind <- sample(nrow(my\_data1), tr\_size)  
data\_tr <- my\_data1[tr\_ind, ] *# training data*  
data\_te <- my\_data1[-tr\_ind, ] *# test data*  
ncol(my\_data1)

## [1] 20

nrow(my\_data1)

## [1] 1649

nrow(data\_tr)

## [1] 1154

nrow(data\_te)

## [1] 495

# Linear model building and statistical diagnosis

set.seed(0)  
  
model <- lm(Life.expectancy ~ ., data = data\_tr)  
summary(model)

##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.3164 -2.0286 0.0263 2.1828 12.0673   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.596e+01 1.076e+00 52.012 < 2e-16 \*\*\*  
## Adult.Mortality -1.606e-02 1.132e-03 -14.192 < 2e-16 \*\*\*  
## infant.deaths 1.061e-01 1.486e-02 7.142 1.64e-12 \*\*\*  
## Alcohol -1.260e-01 3.874e-02 -3.252 0.001181 \*\*   
## Hepatitis.B -1.208e-02 5.174e-03 -2.336 0.019682 \*   
## Measles -1.156e-05 1.311e-05 -0.881 0.378261   
## BMI 3.281e-02 7.058e-03 4.649 3.72e-06 \*\*\*  
## under.five.deaths -7.872e-02 1.070e-02 -7.358 3.57e-13 \*\*\*  
## Polio 1.240e-02 6.128e-03 2.024 0.043191 \*   
## Total.expenditure 2.832e-02 4.782e-02 0.592 0.553864   
## Diphtheria 1.543e-02 7.153e-03 2.156 0.031270 \*   
## HIV.AIDS -4.380e-01 2.149e-02 -20.383 < 2e-16 \*\*\*  
## GDP 5.320e-05 1.089e-05 4.885 1.18e-06 \*\*\*  
## Population -2.122e-09 2.112e-09 -1.005 0.315162   
## thinness..1.19.years -8.094e-03 5.614e-02 -0.144 0.885386   
## thinness.5.9.years -5.641e-02 5.556e-02 -1.015 0.310202   
## Income.composition.of.resources 1.008e+01 1.013e+00 9.949 < 2e-16 \*\*\*  
## Schooling 8.133e-01 7.030e-02 11.569 < 2e-16 \*\*\*  
## Developing -1.288e+00 4.052e-01 -3.178 0.001523 \*\*   
## Expenditure -8.711e-01 2.544e-01 -3.425 0.000637 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.561 on 1134 degrees of freedom  
## Multiple R-squared: 0.8364, Adjusted R-squared: 0.8337   
## F-statistic: 305.2 on 19 and 1134 DF, p-value: < 2.2e-16

Model 1

model\_select <- lm(Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
 under.five.deaths + HIV.AIDS +   
 Income.composition.of.resources + Schooling + Expenditure,  
 data = data\_tr)  
summary(model\_select)

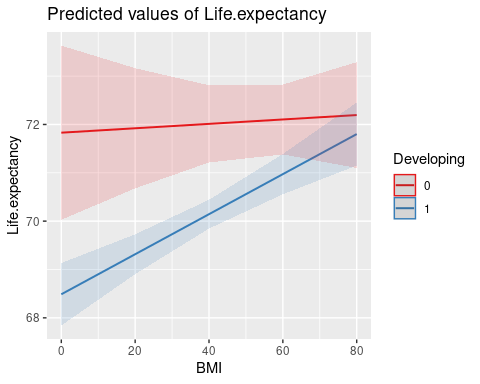
##   
## Call:  
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
## under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling + Expenditure, data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.7705 -2.0389 -0.0376 2.1936 11.4894   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.551772 0.708766 76.967 < 2e-16 \*\*\*  
## BMI 0.035133 0.006719 5.229 2.02e-07 \*\*\*  
## Adult.Mortality -0.017409 0.001141 -15.262 < 2e-16 \*\*\*  
## infant.deaths 0.102012 0.012994 7.851 9.45e-15 \*\*\*  
## under.five.deaths -0.077534 0.009678 -8.012 2.77e-15 \*\*\*  
## HIV.AIDS -0.440200 0.021654 -20.329 < 2e-16 \*\*\*  
## Income.composition.of.resources 11.005533 1.002362 10.980 < 2e-16 \*\*\*  
## Schooling 0.887307 0.066576 13.328 < 2e-16 \*\*\*  
## Expenditure -1.014329 0.255356 -3.972 7.57e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.649 on 1145 degrees of freedom  
## Multiple R-squared: 0.8266, Adjusted R-squared: 0.8254   
## F-statistic: 682.1 on 8 and 1145 DF, p-value: < 2.2e-16

Model 2.a

model\_int <- lm(Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
 under.five.deaths + HIV.AIDS +   
 Income.composition.of.resources + Schooling + Developing +   
 Expenditure + BMI : Developing, data = data\_tr)  
summary(model\_int)

##   
## Call:  
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
## under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling + Developing + Expenditure + BMI:Developing, data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.7697 -2.1114 -0.0309 2.2261 11.5852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 58.737753 1.298487 45.236 < 2e-16 \*\*\*  
## BMI 0.004555 0.016035 0.284 0.776409   
## Adult.Mortality -0.017027 0.001135 -14.997 < 2e-16 \*\*\*  
## infant.deaths 0.107235 0.012956 8.277 3.50e-16 \*\*\*  
## under.five.deaths -0.081345 0.009650 -8.430 < 2e-16 \*\*\*  
## HIV.AIDS -0.440439 0.021496 -20.490 < 2e-16 \*\*\*  
## Income.composition.of.resources 10.555221 1.000057 10.555 < 2e-16 \*\*\*  
## Schooling 0.799797 0.068988 11.593 < 2e-16 \*\*\*  
## Developing -3.342784 0.999664 -3.344 0.000853 \*\*\*  
## Expenditure -1.017215 0.253565 -4.012 6.42e-05 \*\*\*  
## BMI:Developing 0.036868 0.017655 2.088 0.036995 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.621 on 1143 degrees of freedom  
## Multiple R-squared: 0.8295, Adjusted R-squared: 0.828   
## F-statistic: 556 on 10 and 1143 DF, p-value: < 2.2e-16

plot\_model(model\_int, type = "pred", terms = c("BMI", "Developing"))



anova(model\_select, model\_int)

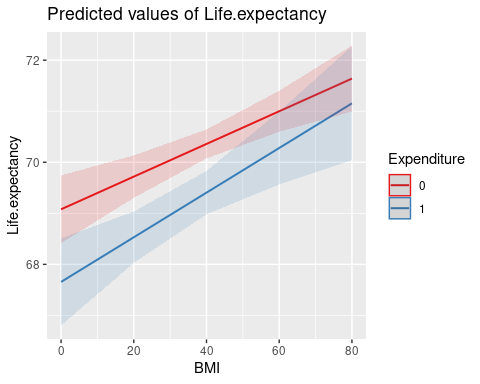
## Analysis of Variance Table  
##   
## Model 1: Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +   
## HIV.AIDS + Income.composition.of.resources + Schooling +   
## Expenditure  
## Model 2: Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +   
## HIV.AIDS + Income.composition.of.resources + Schooling +   
## Developing + Expenditure + BMI:Developing  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1145 15246   
## 2 1143 14990 2 256.26 9.7702 6.205e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Model 2.b

model\_int2 <- lm(Life.expectancy ~ BMI + Adult.Mortality+ infant.deaths +   
 under.five.deaths + HIV.AIDS +   
 Income.composition.of.resources + Schooling + Expenditure +  
 BMI : Expenditure, data = data\_tr)  
summary(model\_int2)

##   
## Call:  
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
## under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling + Expenditure + BMI:Expenditure, data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.8475 -2.0655 -0.0513 2.2334 11.3980   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.771986 0.747599 73.264 < 2e-16 \*\*\*  
## BMI 0.032007 0.007519 4.257 2.24e-05 \*\*\*  
## Adult.Mortality -0.017482 0.001143 -15.289 < 2e-16 \*\*\*  
## infant.deaths 0.102130 0.012995 7.859 8.89e-15 \*\*\*  
## under.five.deaths -0.077544 0.009678 -8.012 2.76e-15 \*\*\*  
## HIV.AIDS -0.439519 0.021668 -20.284 < 2e-16 \*\*\*  
## Income.composition.of.resources 11.024941 1.002643 10.996 < 2e-16 \*\*\*  
## Schooling 0.880288 0.067010 13.137 < 2e-16 \*\*\*  
## Expenditure -1.422313 0.509040 -2.794 0.00529 \*\*   
## BMI:Expenditure 0.011713 0.012642 0.927 0.35438   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.649 on 1144 degrees of freedom  
## Multiple R-squared: 0.8267, Adjusted R-squared: 0.8253   
## F-statistic: 606.4 on 9 and 1144 DF, p-value: < 2.2e-16

plot\_model(model\_int2, type = "pred", terms = c("BMI", "Expenditure"))



anova(model\_select, model\_int2)

## Analysis of Variance Table  
##   
## Model 1: Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +   
## HIV.AIDS + Income.composition.of.resources + Schooling +   
## Expenditure  
## Model 2: Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +   
## HIV.AIDS + Income.composition.of.resources + Schooling +   
## Expenditure + BMI:Expenditure  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 1145 15246   
## 2 1144 15235 1 11.432 0.8584 0.3544

Base on the adjusted R-squared and the P-value of the full model,Using a linear model is appropriate.

A multiple regression was used to study whether the effect of the BMI number on Country’s Developing levels. Results indicated that both BMI and Country’s Developing levels are both associated with the academic performance of the school. The interaction between BMI and Country’s Developing levels is significant. Base on the Anova test, We reject the null hypothesis that the interaction is 0.

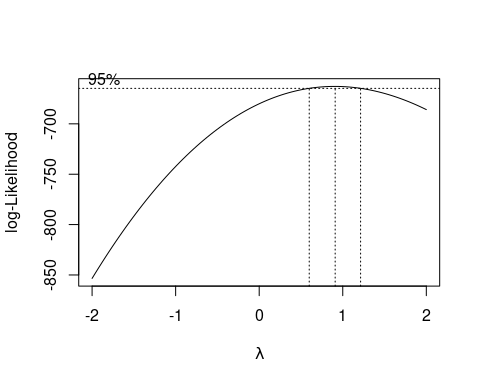
First, we need to do a thorough analysis of the interaction model.

# Model Transformation And Adjustment

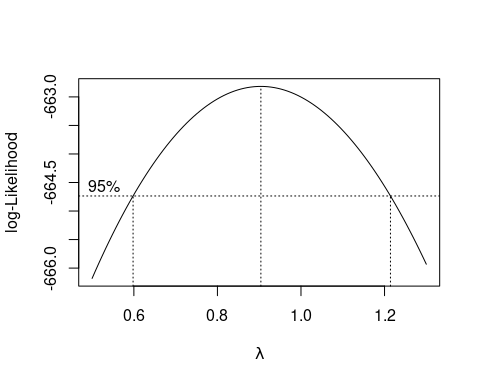
## 1.Box-Cox Transformation

In the previous section we found that there was a problem with the normality of the residuals of the full model, so we tried to solve it using the BOX-COX transform.

boxcox(model\_int, plotit = T)



b <- boxcox(model\_int, plotit = T, lambda = seq(0.5, 1.3, by = 0.01))



I = which(b$y == max(b$y))  
b$x[I]

## [1] 0.9040404

lmod\_trans <- lm(Life.expectancy ^(0.920202) ~ BMI + Adult.Mortality +   
 infant.deaths + under.five.deaths + HIV.AIDS +   
 Income.composition.of.resources + Schooling + Developing +   
 BMI : Developing, data = data\_tr)  
summary(lmod\_trans)

##   
## Call:  
## lm(formula = Life.expectancy^(0.920202) ~ BMI + Adult.Mortality +   
## infant.deaths + under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling + Developing + BMI:Developing, data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.0468 -1.3810 -0.0042 1.4278 7.2789   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 41.6101751 0.8321685 50.002 < 2e-16 \*\*\*  
## BMI 0.0028999 0.0105939 0.274 0.784337   
## Adult.Mortality -0.0111238 0.0007498 -14.836 < 2e-16 \*\*\*  
## infant.deaths 0.0687620 0.0085368 8.055 1.99e-15 \*\*\*  
## under.five.deaths -0.0523101 0.0063603 -8.224 5.28e-16 \*\*\*  
## HIV.AIDS -0.2935748 0.0142014 -20.672 < 2e-16 \*\*\*  
## Income.composition.of.resources 7.2360338 0.6562495 11.026 < 2e-16 \*\*\*  
## Schooling 0.5576471 0.0448294 12.439 < 2e-16 \*\*\*  
## Developing -2.2208643 0.6603160 -3.363 0.000796 \*\*\*  
## BMI:Developing 0.0255797 0.0116594 2.194 0.028442 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.393 on 1144 degrees of freedom  
## Multiple R-squared: 0.8282, Adjusted R-squared: 0.8268   
## F-statistic: 612.7 on 9 and 1144 DF, p-value: < 2.2e-16

dwtest(lmod\_trans)

##   
## Durbin-Watson test  
##   
## data: lmod\_trans  
## DW = 2.0344, p-value = 0.7216  
## alternative hypothesis: true autocorrelation is greater than 0

shapiro.test(lmod\_trans$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: lmod\_trans$residuals  
## W = 0.99123, p-value = 2.28e-06

bptest(lmod\_trans)

##   
## studentized Breusch-Pagan test  
##   
## data: lmod\_trans  
## BP = 105.14, df = 9, p-value < 2.2e-16

Based on the above graph we find that the 95% confidence interval for A contains 1, so we do not see the need to use the BOX-COX transformation.In fact, our model still fails the S-W test after the transformation using the best lambda values, which we believe may be due to problems with the variance of the model residuals.

## 2.Newey-West Adjustments

The presence of heteroskedasticity affects the fit of the linear model, making t-tests and F-tests no longer valid, so in the presence of heteroskedasticity we use heteroskedasticity robust standard errors instead of standard errors. We use white consistent standard errors for hypothesis testing. We use vcovHC() from the sandwich package for this purpose. Also using the NeweyWest() function allows for heteroskedasticity and autocorrelation robustness Newey-West adjustments.

model\_nw <- NeweyWest(model\_int)  
(neweywest <- coeftest(model\_int, vcov = NeweyWest(model\_int)))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 58.7377531 1.3135211 44.7178 < 2.2e-16 \*\*\*  
## BMI 0.0045551 0.0164061 0.2776 0.7813334   
## Adult.Mortality -0.0170271 0.0014028 -12.1381 < 2.2e-16 \*\*\*  
## infant.deaths 0.1072345 0.0156287 6.8614 1.115e-11 \*\*\*  
## under.five.deaths -0.0813451 0.0118983 -6.8367 1.316e-11 \*\*\*  
## HIV.AIDS -0.4404392 0.0278235 -15.8298 < 2.2e-16 \*\*\*  
## Income.composition.of.resources 10.5552210 1.3787175 7.6558 4.072e-14 \*\*\*  
## Schooling 0.7997974 0.0842953 9.4880 < 2.2e-16 \*\*\*  
## Developing -3.3427844 0.9884042 -3.3820 0.0007439 \*\*\*  
## Expenditure -1.0172148 0.2632434 -3.8642 0.0001177 \*\*\*  
## BMI:Developing 0.0368679 0.0178387 2.0667 0.0389834 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(model\_int)

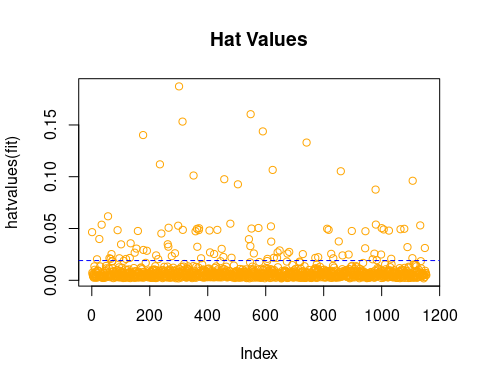
##   
## Call:  
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
## under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling + Developing + Expenditure + BMI:Developing, data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.7697 -2.1114 -0.0309 2.2261 11.5852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 58.737753 1.298487 45.236 < 2e-16 \*\*\*  
## BMI 0.004555 0.016035 0.284 0.776409   
## Adult.Mortality -0.017027 0.001135 -14.997 < 2e-16 \*\*\*  
## infant.deaths 0.107235 0.012956 8.277 3.50e-16 \*\*\*  
## under.five.deaths -0.081345 0.009650 -8.430 < 2e-16 \*\*\*  
## HIV.AIDS -0.440439 0.021496 -20.490 < 2e-16 \*\*\*  
## Income.composition.of.resources 10.555221 1.000057 10.555 < 2e-16 \*\*\*  
## Schooling 0.799797 0.068988 11.593 < 2e-16 \*\*\*  
## Developing -3.342784 0.999664 -3.344 0.000853 \*\*\*  
## Expenditure -1.017215 0.253565 -4.012 6.42e-05 \*\*\*  
## BMI:Developing 0.036868 0.017655 2.088 0.036995 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.621 on 1143 degrees of freedom  
## Multiple R-squared: 0.8295, Adjusted R-squared: 0.828   
## F-statistic: 556 on 10 and 1143 DF, p-value: < 2.2e-16

From the summary table we can see that the robustness estimates differ slightly from the initial estimates, with the variables ‘Polio’,‘Diphtheria’ in the initial estimates changing from significant to insignificant, which confirms the above statement. However, since this adjustment has little effect on either the fitted parameters of the model or the results of the y predictor x significance test j, we also do not intend to use

## 1.Anomaly Detection

### leverage Points

hat\_plot <- **function**(fit){  
 p <- length(coefficients(fit))  
 n <- length(fitted(fit))  
 plot(hatvalues(fit), main = 'Hat Values', col = 'orange')  
 abline(h = 2\*p/n, col = 'blue', lty = 2)  
}  
hat\_plot(model\_int)



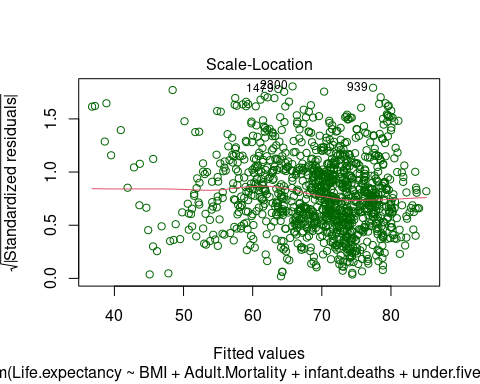
By combining the definition of high leverage points with the diagram above we can see that there are many high leverage points in the model.

### Outliers

check\_outliers(model\_int)

## OK: No outliers detected.  
## - Based on the following method and threshold: cook (0.935).  
## - For variable: (Whole model)

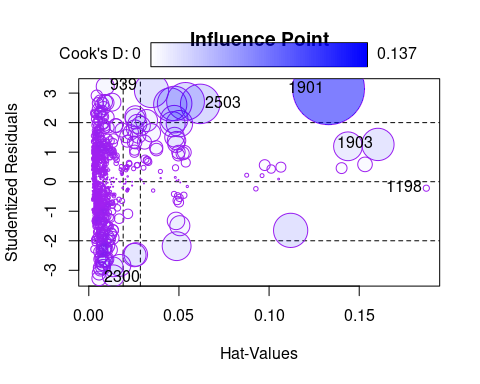
plot(model\_int, which = 3, col = 'darkgreen')



Using the above graph and tests we can obtain that the initial model has no outliers.

### Influential Point

influencePlot(model\_int, id.method = 'identify', main = 'Influence Point',   
 col = 'purple')



## StudRes Hat CookD  
## 2503 2.6349252 0.061790436 0.041353579  
## 2300 -3.2732115 0.005730513 0.005566337  
## 1198 -0.2194212 0.187213505 0.001008991  
## 1903 1.2625972 0.160371377 0.027666288  
## 1901 3.1490650 0.132989259 0.137210659  
## 939 3.2269865 0.009122522 0.008644393

Some anomalies are given in the above graph, but we found that the 1901st sample with the largest Cook distance has a Cook distance value of about 0.1029, which is less than 0.5, and this data sample is large, so we do not think there are strong influence points that need to be removed from this model.

## Analysis of Gaussian-Markov Assumptions

### Zero-mean Assumption

mean(model\_int$residuals)

## [1] -4.122193e-16

Based on the above calculations, the model residuals are very close to 0.

### Homoskedasticity Assumption

bptest(model\_int)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_int  
## BP = 91.973, df = 10, p-value = 2.174e-15

bptest(model\_int, studentize = F)

##   
## Breusch-Pagan test  
##   
## data: model\_int  
## BP = 121.77, df = 10, p-value < 2.2e-16

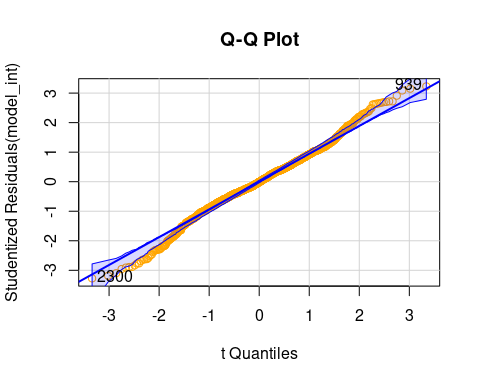
We found that although the p-values did not differ they were all less than 0.05, indicating that there was strong heteroskedasticity in the model. However, the BP values with studentisation removed increased, suggesting that studentisation played a role in correcting for heteroskedasticity, but not significantly in this case.

### Normality Assumption

shapiro.test(model\_int$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model\_int$residuals  
## W = 0.99133, p-value = 2.584e-06

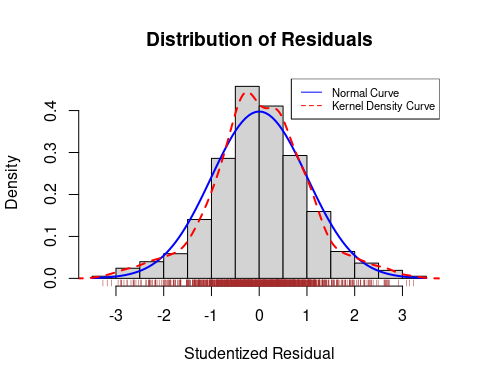
qqPlot(model\_int, labels = row.names(df), id.method = 'identify', simulate = T,  
 main = 'Q-Q Plot', col = 'orange')



## 2300 939   
## 135 827

In the Q-Q plot above, the blue shaded area is the 95% confidence interval and the two outlier sample points that were detected, for the 1901st and 2300th samples.

residplot <- **function**(model\_int, nbreaks = 10){  
 z <- rstudent(model\_int)  
 hist(z,breaks = nbreaks, freq = F,  
 xlab = 'Studentized Residual',  
 main = 'Distribution of Residuals')  
 rug(jitter(z), col = 'brown')  
 curve(dnorm(x, mean = mean(z), sd = sd(z)), add = T, col = 'blue', lwd = 2)  
 lines(density(z)$x, density(z)$y, col = "red", lwd = 2, lty = 2)  
 legend('topright', legend = c('Normal Curve', 'Kernel Density Curve'),  
 lty = 1:2, col = c('blue', 'red'), cex = .7)  
}  
residplot(model\_int)



We can see from the residual distribution graph that the model residuals are almost completely unbiased. This is one of the reasons why subsequently when we used the BOX-COX variation to calculate the lambda we found that its confidence interval contained 1, i.e. the BOX-COX transformation was not necessary. In addition to this the problem of heteroskedasticity can also have an impact on the effectiveness of the BOX-COX transformation.

From the graphs above and the results of the tests we can conclude that the initial model residuals do not obey normality, but rather suffer from some heavy tails.

### Linearity Assumption

We would have liked to use a deviation residual plot for this test, but the model has too many predictors and a large sample, and the RMD does not have enough computing power to give results. At the end of this section, we will use check\_model() to find out about linearity.

### Randomness Assumption

dwtest(model\_int)

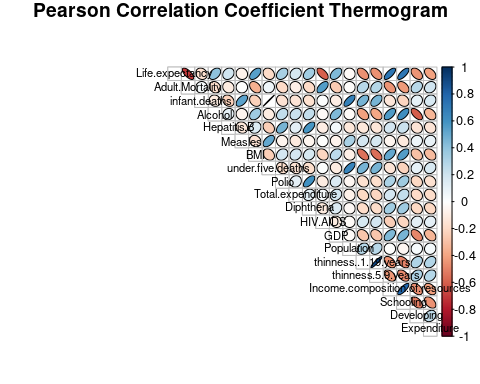
##   
## Durbin-Watson test  
##   
## data: model\_int  
## DW = 2.032, p-value = 0.7081  
## alternative hypothesis: true autocorrelation is greater than 0

From the p-values of the above results we can conclude that there is no first order autocorrelation problem with the model.

### No Multicollinearity Assumption

First we can take a cursory look at the two-by-two correlation between the variables using a thermogram of the Pearson correlation coefficient matrix.

M = cor(my\_data1)  
corrplot(M, method = 'ellipse', type = 'upper', tl.col = 'black', tl.pos= 'd',   
 tl.cex = 0.7, show.legend = T, outline = T,   
 title = 'Pearson Correlation Coefficient Thermogram',   
 mar = c(0, 0, 1, 0))



However, Pearson’s correlation coefficient can only show the correlation between two variables. In practical problems there may be problems with correlations between more than one variable, so for a further and clearer view we introduce the variance inflation factor.

alias(model\_int)

## Model :  
## Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +   
## HIV.AIDS + Income.composition.of.resources + Schooling +   
## Developing + Expenditure + BMI:Developing

The above checks revealed that none of the predictors in the data had a large number of identical data, leading to problems where parameters could not be fitted or vif could not be calculated.

pander(vif(model\_int), caption = 'Vif of Full Model')

*Table continues below*

| BMI | Adult.Mortality | infant.deaths | under.five.deaths | HIV.AIDS |
| --- | --- | --- | --- | --- |
| 8.859 | 1.77 | 188.3 | 188.8 | 1.452 |

*Table continues below*

| Income.composition.of.resources | Schooling | Developing | Expenditure |
| --- | --- | --- | --- |
| 2.907 | 3.416 | 11.05 | 1.293 |

| BMI:Developing |
| --- |
| 12.88 |

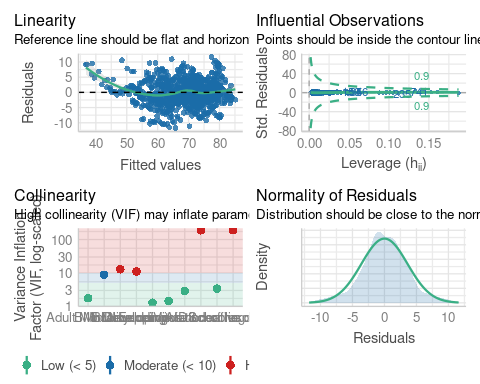
Using the above graphs we find that several predictors of ‘infant.deaths’, ‘per centage.expenditure’, ‘under.five.deaths’, ‘GDP’ have VIFs greater than 10 and their presence leads to serious multicollinearity problems.

## 3.Model overview

Finally, let’s look at the statistical diagnosis of the full model as a whole.

check\_model(model\_int, verbose = T,   
 check = c('outliers', 'vif', 'normality', 'linearity'))

## Model has interaction terms. VIFs might be inflated.  
## You may check multicollinearity among predictors of a model without  
## interaction terms.



In this section we find that the pass test results for the full model, although fair overall, suffer mainly from multicollinearity, heteroskedasticity and non-normality. In the next section we will try to address these problems using the methods we have learned.

# Model Variables Selection

## 1.AIC Selection

step(model\_int)

## Start: AIC=2981.03  
## Life.expectancy ~ BMI + Adult.Mortality + infant.deaths + under.five.deaths +   
## HIV.AIDS + Income.composition.of.resources + Schooling +   
## Developing + Expenditure + BMI:Developing  
##   
## Df Sum of Sq RSS AIC  
## <none> 14990 2981.0  
## - BMI:Developing 1 57.2 15047 2983.4  
## - Expenditure 1 211.1 15201 2995.2  
## - infant.deaths 1 898.4 15888 3046.2  
## - under.five.deaths 1 932.0 15922 3048.6  
## - Income.composition.of.resources 1 1461.0 16451 3086.3  
## - Schooling 1 1762.7 16753 3107.3  
## - Adult.Mortality 1 2949.6 17940 3186.3  
## - HIV.AIDS 1 5505.9 20496 3340.0

##   
## Call:  
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
## under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling + Developing + Expenditure + BMI:Developing, data = data\_tr)  
##   
## Coefficients:  
## (Intercept) BMI   
## 58.737753 0.004555   
## Adult.Mortality infant.deaths   
## -0.017027 0.107235   
## under.five.deaths HIV.AIDS   
## -0.081345 -0.440439   
## Income.composition.of.resources Schooling   
## 10.555221 0.799797   
## Developing Expenditure   
## -3.342784 -1.017215   
## BMI:Developing   
## 0.036868

*# Find model with lowest AIC*  
lmod\_AIC\_B <- lm(Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
 under.five.deaths + HIV.AIDS +   
 Income.composition.of.resources + Schooling + Developing +   
 BMI:Developing, data = data\_tr) *# AIC selected model*  
sum\_AIC\_B <- summary(lmod\_AIC\_B)  
sum\_AIC\_B

##   
## Call:  
## lm(formula = Life.expectancy ~ BMI + Adult.Mortality + infant.deaths +   
## under.five.deaths + HIV.AIDS + Income.composition.of.resources +   
## Schooling + Developing + BMI:Developing, data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.1472 -2.1220 -0.0016 2.1714 10.9449   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 57.472249 1.267867 45.330 < 2e-16 \*\*\*  
## BMI 0.004386 0.016141 0.272 0.785878   
## Adult.Mortality -0.016899 0.001142 -14.793 < 2e-16 \*\*\*  
## infant.deaths 0.103435 0.013006 7.953 4.36e-15 \*\*\*  
## under.five.deaths -0.078702 0.009690 -8.122 1.18e-15 \*\*\*  
## HIV.AIDS -0.440082 0.021637 -20.339 < 2e-16 \*\*\*  
## Income.composition.of.resources 11.020416 0.999842 11.022 < 2e-16 \*\*\*  
## Schooling 0.849761 0.068301 12.441 < 2e-16 \*\*\*  
## Developing -3.422544 1.006038 -3.402 0.000692 \*\*\*  
## BMI:Developing 0.038866 0.017764 2.188 0.028877 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.645 on 1144 degrees of freedom  
## Multiple R-squared: 0.8271, Adjusted R-squared: 0.8257   
## F-statistic: 608 on 9 and 1144 DF, p-value: < 2.2e-16

From the summary we can find that the model selected by the backward iterative AIC method, most certainly all predictors are statistically significant, but the adjusted R-squared does not change much compared to the full model, and we will subsequently judge whether this model should be used by the model’s prediction error perspective

## 2.BIC Selection

set.seed(0)  
fit\_null <- lm(Life.expectancy ~ 1, data\_tr)  
step(fit\_null, scope = list(lower = fit\_null, upper = model\_int),   
 direction = "both", criterion = "BIC")

## Start: AIC=5002.38  
## Life.expectancy ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + Schooling 1 46348 41563 4139.9  
## + Income.composition.of.resources 1 46324 41586 4140.6  
## + Adult.Mortality 1 42969 44941 4230.1  
## + HIV.AIDS 1 30508 57403 4512.5  
## + BMI 1 24795 63116 4622.0  
## + Developing 1 17741 70169 4744.3  
## + Expenditure 1 12653 75258 4825.1  
## + under.five.deaths 1 3146 84765 4962.3  
## + infant.deaths 1 2460 85450 4971.6  
## <none> 87911 5002.4  
##   
## Step: AIC=4139.91  
## Life.expectancy ~ Schooling  
##   
## Df Sum of Sq RSS AIC  
## + HIV.AIDS 1 17966 23597 3488.7  
## + Adult.Mortality 1 16736 24827 3547.3  
## + Income.composition.of.resources 1 5400 36164 3981.3  
## + BMI 1 2063 39500 4083.2  
## + Developing 1 555 41008 4126.4  
## + Expenditure 1 288 41275 4133.9  
## <none> 41563 4139.9  
## + under.five.deaths 1 33 41530 4141.0  
## + infant.deaths 1 2 41561 4141.9  
## - Schooling 1 46348 87911 5002.4  
##   
## Step: AIC=3488.65  
## Life.expectancy ~ Schooling + HIV.AIDS  
##   
## Df Sum of Sq RSS AIC  
## + Adult.Mortality 1 4770 18827 3230.1  
## + Income.composition.of.resources 1 3110 20487 3327.5  
## + BMI 1 984 22613 3441.5  
## + Developing 1 395 23202 3471.2  
## + Expenditure 1 340 23258 3473.9  
## + under.five.deaths 1 98 23499 3485.8  
## + infant.deaths 1 47 23550 3488.3  
## <none> 23597 3488.7  
## - HIV.AIDS 1 17966 41563 4139.9  
## - Schooling 1 33805 57403 4512.5  
##   
## Step: AIC=3230.06  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality  
##   
## Df Sum of Sq RSS AIC  
## + Income.composition.of.resources 1 2088.5 16739 3096.4  
## + BMI 1 623.1 18204 3193.2  
## + Expenditure 1 380.7 18447 3208.5  
## + Developing 1 231.8 18596 3217.8  
## + under.five.deaths 1 105.0 18722 3225.6  
## + infant.deaths 1 63.3 18764 3228.2  
## <none> 18827 3230.1  
## - Adult.Mortality 1 4769.9 23597 3488.7  
## - HIV.AIDS 1 5999.7 24827 3547.3  
## - Schooling 1 20937.3 39765 4090.9  
##   
## Step: AIC=3096.37  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources  
##   
## Df Sum of Sq RSS AIC  
## + BMI 1 425.0 16314 3068.7  
## + Expenditure 1 207.4 16532 3084.0  
## + under.five.deaths 1 157.3 16582 3087.5  
## + Developing 1 124.6 16614 3089.7  
## + infant.deaths 1 109.6 16629 3090.8  
## <none> 16739 3096.4  
## - Income.composition.of.resources 1 2088.5 18827 3230.1  
## - Adult.Mortality 1 3748.0 20487 3327.5  
## - Schooling 1 3916.8 20656 3337.0  
## - HIV.AIDS 1 5740.0 22479 3434.6  
##   
## Step: AIC=3068.7  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +   
## BMI  
##   
## Df Sum of Sq RSS AIC  
## + Expenditure 1 170.0 16144 3058.6  
## + Developing 1 122.2 16192 3062.0  
## + under.five.deaths 1 91.7 16222 3064.2  
## + infant.deaths 1 55.3 16259 3066.8  
## <none> 16314 3068.7  
## - BMI 1 425.0 16739 3096.4  
## - Income.composition.of.resources 1 1890.3 18204 3193.2  
## - Schooling 1 2911.2 19225 3256.2  
## - Adult.Mortality 1 3528.9 19843 3292.7  
## - HIV.AIDS 1 5614.7 21929 3408.0  
##   
## Step: AIC=3058.61  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +   
## BMI + Expenditure  
##   
## Df Sum of Sq RSS AIC  
## + Developing 1 125.6 16018 3051.6  
## + under.five.deaths 1 77.0 16067 3055.1  
## + infant.deaths 1 43.1 16101 3057.5  
## <none> 16144 3058.6  
## - Expenditure 1 170.0 16314 3068.7  
## - BMI 1 387.5 16532 3084.0  
## - Income.composition.of.resources 1 1749.9 17894 3175.4  
## - Schooling 1 2533.9 18678 3224.9  
## - Adult.Mortality 1 3583.4 19727 3287.9  
## - HIV.AIDS 1 5629.4 21773 3401.8  
##   
## Step: AIC=3051.6  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +   
## BMI + Expenditure + Developing  
##   
## Df Sum of Sq RSS AIC  
## + under.five.deaths 1 79.3 15939 3047.9  
## + BMI:Developing 1 59.0 15959 3049.3  
## + infant.deaths 1 43.7 15975 3050.4  
## <none> 16018 3051.6  
## - Developing 1 125.6 16144 3058.6  
## - Expenditure 1 173.4 16192 3062.0  
## - BMI 1 384.8 16403 3077.0  
## - Income.composition.of.resources 1 1653.1 17672 3162.9  
## - Schooling 1 2080.8 18099 3190.5  
## - Adult.Mortality 1 3503.6 19522 3277.9  
## - HIV.AIDS 1 5654.8 21673 3398.5  
##   
## Step: AIC=3047.87  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +   
## BMI + Expenditure + Developing + under.five.deaths  
##   
## Df Sum of Sq RSS AIC  
## + infant.deaths 1 891.9 15047 2983.4  
## + BMI:Developing 1 50.7 15888 3046.2  
## <none> 15939 3047.9  
## - under.five.deaths 1 79.3 16018 3051.6  
## - Developing 1 127.9 16067 3055.1  
## - Expenditure 1 158.4 16097 3057.3  
## - BMI 1 328.3 16267 3069.4  
## - Income.composition.of.resources 1 1695.9 17635 3162.6  
## - Schooling 1 1952.6 17892 3179.2  
## - Adult.Mortality 1 3509.8 19449 3275.5  
## - HIV.AIDS 1 5687.4 21626 3398.0  
##   
## Step: AIC=2983.42  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +   
## BMI + Expenditure + Developing + under.five.deaths + infant.deaths  
##   
## Df Sum of Sq RSS AIC  
## + BMI:Developing 1 57.2 14990 2981.0  
## <none> 15047 2983.4  
## - Developing 1 199.1 15246 2996.6  
## - Expenditure 1 217.5 15265 2998.0  
## - BMI 1 361.5 15409 3008.8  
## - infant.deaths 1 891.9 15939 3047.9  
## - under.five.deaths 1 927.5 15975 3050.4  
## - Income.composition.of.resources 1 1484.3 16532 3090.0  
## - Schooling 1 1864.2 16911 3116.2  
## - Adult.Mortality 1 2988.3 18036 3190.5  
## - HIV.AIDS 1 5528.9 20576 3342.6  
##   
## Step: AIC=2981.03  
## Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +   
## BMI + Expenditure + Developing + under.five.deaths + infant.deaths +   
## BMI:Developing  
##   
## Df Sum of Sq RSS AIC  
## <none> 14990 2981.0  
## - BMI:Developing 1 57.2 15047 2983.4  
## - Expenditure 1 211.1 15201 2995.2  
## - infant.deaths 1 898.4 15888 3046.2  
## - under.five.deaths 1 932.0 15922 3048.6  
## - Income.composition.of.resources 1 1461.0 16451 3086.3  
## - Schooling 1 1762.7 16753 3107.3  
## - Adult.Mortality 1 2949.6 17940 3186.3  
## - HIV.AIDS 1 5505.9 20496 3340.0

##   
## Call:  
## lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +   
## Income.composition.of.resources + BMI + Expenditure + Developing +   
## under.five.deaths + infant.deaths + BMI:Developing, data = data\_tr)  
##   
## Coefficients:  
## (Intercept) Schooling   
## 58.737753 0.799797   
## HIV.AIDS Adult.Mortality   
## -0.440439 -0.017027   
## Income.composition.of.resources BMI   
## 10.555221 0.004555   
## Expenditure Developing   
## -1.017215 -3.342784   
## under.five.deaths infant.deaths   
## -0.081345 0.107235   
## BMI:Developing   
## 0.036868

lmod\_BIC\_BO <- lm(Life.expectancy ~ BMI:Developing + infant.deaths +   
 under.five.deaths + Income.composition.of.resources +   
 Schooling + Adult.Mortality + HIV.AIDS, data = data\_tr)   
 *# BIC selected model*  
sum\_BIC\_BO <- summary(lmod\_BIC\_BO)  
sum\_BIC\_BO

##   
## Call:  
## lm(formula = Life.expectancy ~ BMI:Developing + infant.deaths +   
## under.five.deaths + Income.composition.of.resources + Schooling +   
## Adult.Mortality + HIV.AIDS, data = data\_tr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.3251 -2.1229 0.0041 2.2902 11.4881   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 53.042903 0.657226 80.707 < 2e-16 \*\*\*  
## infant.deaths 0.095150 0.013206 7.205 1.05e-12 \*\*\*  
## under.five.deaths -0.073116 0.009847 -7.425 2.19e-13 \*\*\*  
## Income.composition.of.resources 12.043008 1.008314 11.944 < 2e-16 \*\*\*  
## Schooling 1.030716 0.064291 16.032 < 2e-16 \*\*\*  
## Adult.Mortality -0.017780 0.001158 -15.360 < 2e-16 \*\*\*  
## HIV.AIDS -0.443898 0.022056 -20.126 < 2e-16 \*\*\*  
## BMI:Developing 0.005921 0.005109 1.159 0.247   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.717 on 1146 degrees of freedom  
## Multiple R-squared: 0.8199, Adjusted R-squared: 0.8188   
## F-statistic: 745.2 on 7 and 1146 DF, p-value: < 2.2e-16

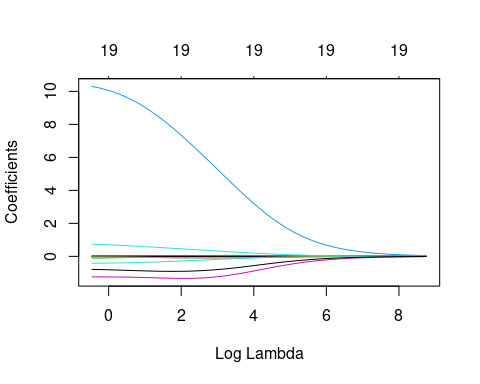
From the summary we can find that the model chosen by the backward iterative BIC method, most certainly all predictors are statistically significant, but the adjusted R-squared is even lower than the full model, and subsequently we will judge whether this model should be used by the model’s prediction error perspective.

## 3.Selection ideas for other model selection methods

In the statistical diagnosis section of the model, we find that the full model suffers from multicollinearity and heteroskedasticity. GLS estimation is usually used when the m-model error term does not satisfy the “spherical perturbation assumption” (i.e. homoskedasticity assumption and no autocorrelation assumption in the G-M assumption). Ridge regression, lasso regression and adaptive lasso regression are all methods of constraining the fitted parameters by adding penalty factors. We will try each of these below.

## 4.Ridge Selection

set.seed(0)  
x\_tr <- as.matrix(data\_tr[, c(2 : ncol(data\_tr))])  
y\_tr <- as.matrix(data\_tr[, 1])  
x\_te <- as.matrix(data\_te[, c(2 : ncol(data\_te))])  
y\_te <- as.matrix(data\_te[, 1])  
set.seed(0)  
ridge <- glmnet(x = x\_tr, y = y\_tr, alpha = 0)  
plot(ridge, xvar = 'lambda')



ridge\_cv <- cv.glmnet(x = x\_tr, y = y\_tr, type.measure = 'mse', nfold = 10,  
 alpha = 0)

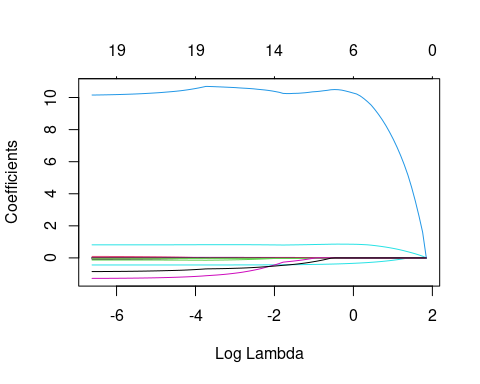
ridge\_cv$lambda.min

## [1] 0.6337393

best\_ridge <- coef(ridge\_cv, s = ridge\_cv$lambda.min)

## 5.Lasso Selection

set.seed(0)  
lasso <- glmnet(x = x\_tr, y = y\_tr, alpha = 1)  
plot(lasso, xvar = 'lambda')

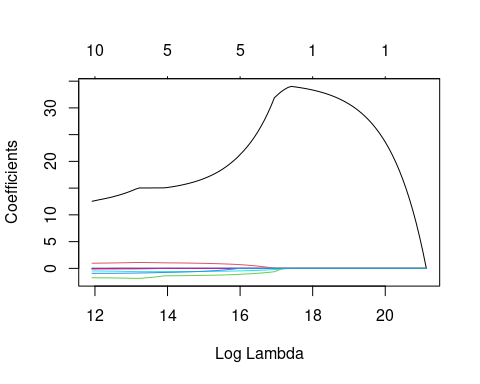


lasso\_cv <- cv.glmnet(x = x\_tr, y = y\_tr, type.measure = 'mse', nfold = 10,   
 alpha = 1, keep = T)  
lasso\_cv$lambda.min

## [1] 0.002558417

## 6.Adaptive Lasso Selection

set.seed(0)  
alasso <- glmnet(x = x\_tr, y = y\_tr, alpha = 1,   
 penalty.factor = 1/abs(best\_ridge[-1]))  
plot(alasso, xvar = 'lambda')



alasso\_cv <- cv.glmnet(x = x\_tr, y = y\_tr, type.measure = 'mse', nfold = 10,  
 alpha = 1, penalty.factor = 1/abs(best\_ridge[-1]),  
 keep = T)

alasso\_cv$lambda.min

## [1] 150413.6

## 7.Error Comparison And Confirmation of Final Model

Next we will calculate the prediction error of each model in the training set:

result\_full <- predict(model\_int, newdata = data\_te, interval = 'prediction')   
(err\_full <- mean((data\_te$Life.expectancy - result\_full) ^2))

## [1] 47.1422

result\_aic <- predict(lmod\_AIC\_B, newdata = data\_te, interval = 'prediction')   
(err\_aic <- mean((data\_te$Life.expectancy - result\_aic) ^2))

## [1] 48.02168

result\_bic <- predict(lmod\_BIC\_BO, newdata = data\_te, interval = 'prediction')   
(err\_bic <- mean((data\_te$Life.expectancy - result\_bic) ^2))

## [1] 49.85988

result\_ridge <- predict(ridge\_cv, newx = x\_te, interval = 'prediction')   
(err\_ridge <- mean((y\_te - result\_ridge) ^2))

## [1] 14.42922

result\_la <- predict(lasso\_cv, newx = x\_te, interval = 'prediction')   
(err\_la <- mean((y\_te - result\_la) ^2))

## [1] 13.48764

result\_adala <- predict(alasso\_cv, newx = x\_te, interval = 'prediction')   
(err\_adala <- mean((y\_te - result\_adala) ^2))

## [1] 15.42448

which.min(data.frame(err\_full, err\_aic, err\_bic, err\_ridge, err\_la, err\_adala))

## err\_la   
## 5

From the above results we can see that the model selected using the 10-fold lasso method has the smallest test error and a significant reduction compared to the original model, so we will finally choose this model.

(best\_alasso\_coef <- coef(alasso\_cv, s = alasso\_cv$lambda.min))

## 20 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 54.350598539  
## Adult.Mortality -0.012711660  
## infant.deaths .   
## Alcohol -0.136282144  
## Hepatitis.B .   
## Measles .   
## BMI 0.022720708  
## under.five.deaths .   
## Polio .   
## Total.expenditure .   
## Diphtheria 0.004019068  
## HIV.AIDS -0.487274911  
## GDP .   
## Population .   
## thinness..1.19.years .   
## thinness.5.9.years -0.037510959  
## Income.composition.of.resources 12.544671846  
## Schooling 0.957734593  
## Developing -1.753032977  
## Expenditure -0.921002466

So our final model will be:

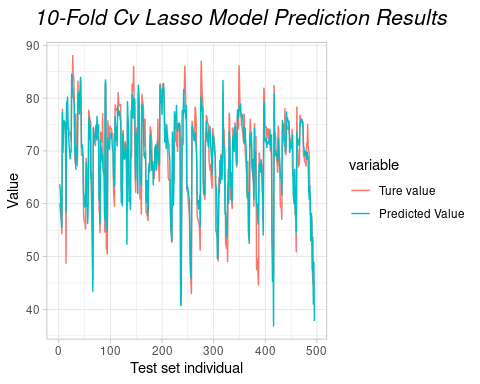
$$Life.expectancy=53.429596073-0.012711660\*Adult.Mortality-0.136282144\*Alcohol+0.022720708\*BMI+\\\\0.004019068\*Diphtheria-0.487274911\*HIV.AIDS -0.037510959\*thinness.5.9.years+\\\\12.544671846\*Income.composition.of.resources + 0.957734593\*Schooling\\\\-1.753032977\*Developing - 0.921002466 \* Expenditure$$

# Model prediction

In this section we will use our selected 10-fold lasso model to make predictions and compare them with the true values, by way of icons to see the predictions.

x\_gr <- 1:495  
y\_pred <- predict(lasso\_cv, x\_te)  
df <- data.frame(x\_gr, y\_te, y\_pred)  
names(df) <- c('x\_gr', 'Ture value', 'Predicted Value')  
df\_long <- melt(df, id.vars = 'x\_gr')  
P <- ggplot(df\_long, aes(x\_gr, value, col = variable)) +  
 geom\_xspline() + labs(x = 'Test set individual', y = 'Value') + theme\_light()  
grid.arrange(textGrob('10-Fold Cv Lasso Model Prediction Results',  
 gp = gpar(fontsize =2\*8, fontface ='italic')),   
 P,   
 heights=c(0.1,1))

## Warning: Using the `size` aesthetic in this geom was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` in the `default\_aes` field and elsewhere instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



sst <- sum((y\_te - mean(y\_te))^2)  
sse <- sum((y\_pred - y\_te)^2)  
rsq <- 1 - sse/sst  
rsq

## [1] 0.8307413

As we can see from the graph above, the predicted values are close to the true values, which means that the model is successful in its predictions.

Code Book for our Data:

[GPH GU 2353 3353\_Codebook for Group A Plus Data\_FinalProject](https://docs.google.com/spreadsheets/d/1YFq9jB07XF_jFaI-2mznqwaUGsg2Pm6Hza0TqOoL36E/edit?usp=share_link)

Open Source HTTPs links:

<https://github.com/yangy37/GPH2353-Project.git>.