GR5205-LINEAR REGRESSION MODEL:FINAL PROJECT

ZIJIE XIA, UNI:ZX2276

CONTENTS

LIST OF FIGURES

1 PART I

1.1 Introduction

The goal of the section is to find the statistically significant relationship between democracy rate and infant death rate (without controlling for life expectancy). Before building multiple linear regression models and running classic hypothesis testing procedures, data will be explored.

Analyse the data frame at first. After dropping all NAs of the data frame, 165 records are left. Since democracy rate is the mean of five measures, that is electoral process and pluralism, function of government, political participation, political culture, civil liberties, these measures are useless in the model and they can be deleted. Meanwhile, regime type is determined by democracy rate and it will not be used therefore. By checking correlation matrix, we find male life expectancy, female life expectancy and life expectancy are extremely highly correlated. So, male and female life expectancy can be ignored. The left variables are what we should care about.

In this section, democracy rate is supposed as the response variable. We can see some properties of the variable. By Kolmogorov-Smirnov test, the distribution of democracy rate is not normal under 95 percent confidence interval. Therefore, we need to remedy when doing regression. Also, by QQ-Plot and Box-Plot, we find the distribution of response variable is short-tailed.

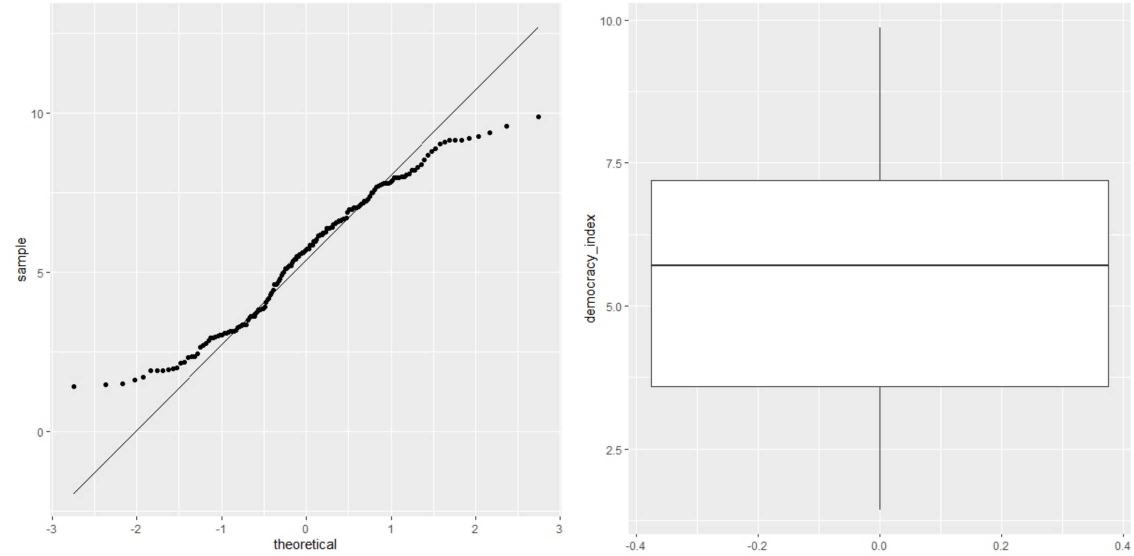

Figure 1: QQ-Plot and Box-Plot about the response variable

1.2 Statistical Model

After analysis and trials, the model and its results are as followed:

```
> summary(model1_9)

Call:
lm(formula = democracy_index ~ region + health_spend_pct_gdp +
    gdpPPP_percap + land_area + coastline + infant_mortality_rate,
    data = vars3, weights = cal.weights)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
−4.9499 −0.7902  0.0580  0.8249  2.9936

```

```
Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                               4.825e+00  5.510e-01   8.757 4.07e-15 ***
regionAsia                               -1.167e-02  7.745e-01  -0.015 0.988000
regionCentral America and the Caribbean  -1.195e-01  5.071e-01  -0.236 0.814031
regionEurasia                            -1.554e+00  5.571e-01  -2.790 0.005962 **
regionEurope                              6.325e-01  4.548e-01   1.391 0.166372
regionMiddle East                        -2.505e+00  5.641e-01  -4.441 1.74e-05 ***
regionNorth America                       4.664e-02  1.006e+00   0.046 0.963073
regionOceania                             1.762e+00  6.205e-01   2.840 0.005149 **
regionSouth America                       1.011e+00  4.721e-01   2.141 0.033931 *
regionSouth Asia                          9.196e-01  5.733e-01   1.604 0.110819
regionSoutheast Asia                     -2.977e-01  5.136e-01  -0.580 0.563116
health_spend_pct_gdp                      1.683e-01  4.693e-02   3.587 0.000453 ***
gdpPPP_percap                             2.574e-05  6.136e-06   4.194 4.68e-05 ***
land_area                                -1.911e-07  7.121e-08  -2.683 0.008118 **
coastline                                 1.673e-05  5.542e-06   3.019 0.002981 **
infant_mortality_rate                    -3.305e-02  9.040e-03  -3.656 0.000354 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 1.314 on 149 degrees of freedom
Multiple R-squared:  0.6952,  Adjusted R-squared:  0.6645
F-statistic: 22.66 on 15 and 149 DF,  p-value: < 2.2e-16

> anova(model1_9)
Analysis of Variance Table

Response: democracy_index
                       Df Sum Sq Mean Sq F value    Pr(>F)
region                 10 448.06  44.806 25.9493 < 2.2e-16 ***
health_spend_pct_gdp    1  35.82  35.818 20.7435 1.084e-05 ***
gdpPPP_percap           1  55.51  55.510 32.1481 7.186e-08 ***
land_area               1   7.00   7.001  4.0546 0.0458521 *
coastline               1  17.41  17.405 10.0802 0.0018214 **
infant_mortality_rate   1  23.08  23.084 13.3689 0.0003541 ***
Residuals             149 257.28   1.727
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1
>
```

Algorithm 1: Result of the model

Visualize model 1.9 as Figure 2.

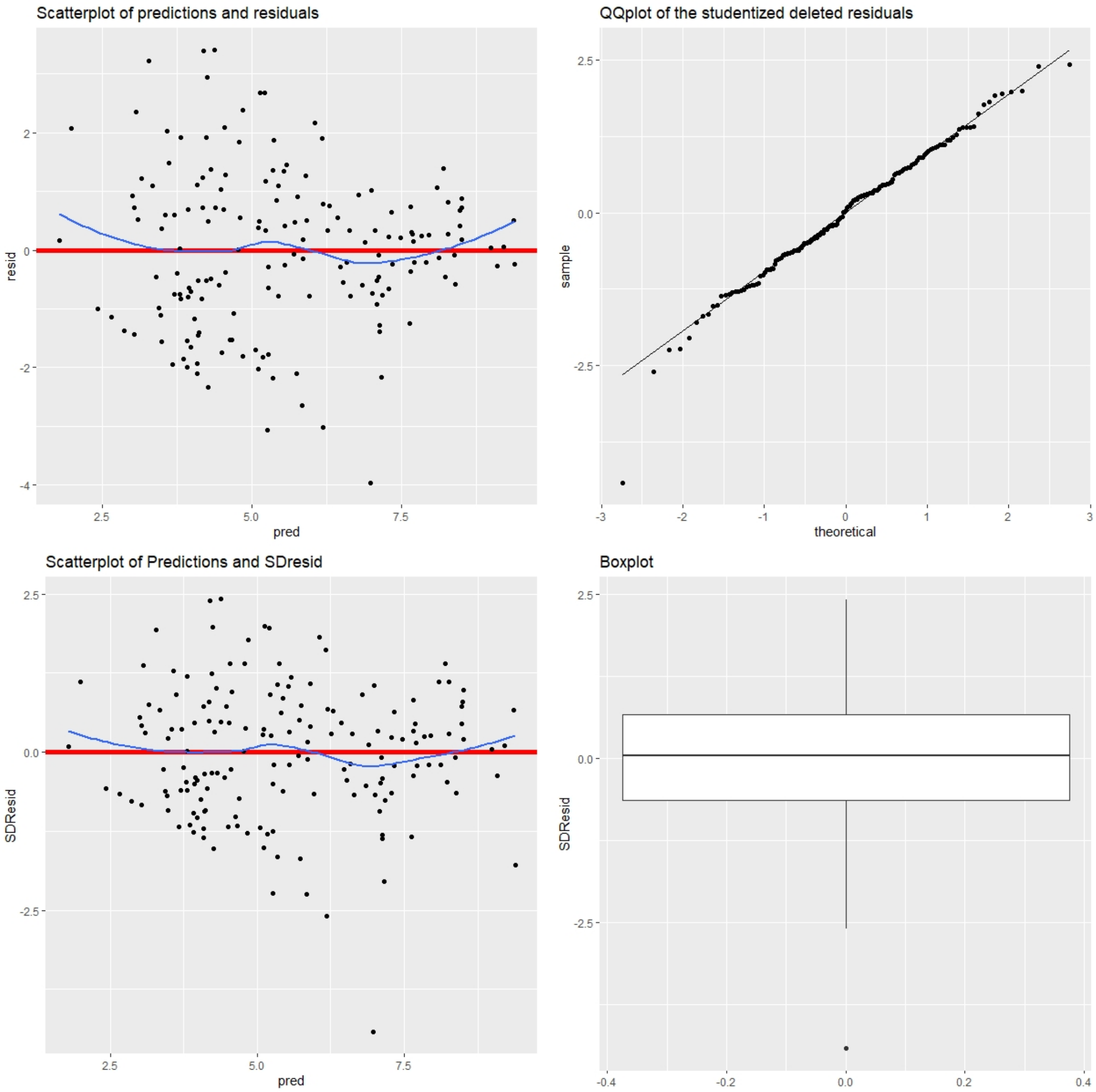

Figure 2: Properties of model 1.9

Variables are region, infant mortality rate, health spend pct gdp, gdpPPP percap, land area, coastline. Multiple R-squared is 0.695 and Adjusted R-squared is 0.6645, which means the performance of the model is not bad. Hypothesis is tested on samples without outliers and there is no interactions. It is not necessary to merge dummies in region variable here considering the goal. But it will be important in next section. Transformation of variables will be tried in the next section.

1.3 Research Question

As shown in t-test, p-value of infant mortality rate is 0.0003541. And in f-test, on condition of existence of other variables, p-value of infant mortality rate is 0.0003541. So, reject null hypothesis, that is, there is statistically significant relationship between democracy rate and infant death rate.

1.4 Appendix

1.4.1 *Model (including Diagnostics)*

Firstly, choose variables using forward selection with AIC. The covariates are region, infant mortality rate, health spend pct gdp, gdpPPP percap, land area, coastline, death rate, roadways, refined petrol consumption, birth rate and airports.

```
> summary(model1_1)

Call:
lm(formula = democracy_index ~ region + infant_mortality_rate +
    health_spend_pct_gdp + gdpPPP_percap + land_area + coastline +
    death_rate + roadways + refined_petrol_consumption + birth_rate +
    airports, data = vars3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2468 -0.7994  0.0444  0.7418  3.4849

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            3.414e+00  8.737e-01   3.908 0.000143 ***
regionAsia                             1.541e+00  1.142e+00   1.349 0.179418
regionCentral America and the Caribbean 3.778e-01  5.251e-01   0.720 0.472961
regionEurasia                         -1.455e+00  5.261e-01  -2.765 0.006429 **
regionEurope                           4.246e-01  5.444e-01   0.780 0.436737
regionMiddle East                     -1.755e+00  5.122e-01  -3.426 0.000798 ***
regionNorth America                    3.951e-01  1.461e+00   0.270 0.787167
regionOceania                          1.906e+00  7.572e-01   2.517 0.012926 *
regionSouth America                    1.248e+00  5.411e-01   2.306 0.022545 *
regionSouth Asia                       9.468e-01  6.409e-01   1.477 0.141768
regionSoutheast Asia                   1.447e-01  5.211e-01   0.278 0.781654
infant_mortality_rate                 -5.817e-02  1.295e-02  -4.492 1.44e-05 ***
health_spend_pct_gdp                   1.225e-01  5.403e-02   2.267 0.024901 *
gdpPPP_percap                          2.737e-05  7.270e-06   3.765 0.000242 ***
land_area                             -3.569e-07  9.196e-08  -3.882 0.000157 ***
coastline                              2.225e-05  9.813e-06   2.267 0.024882 *
death_rate                             1.399e-01  6.019e-02   2.325 0.021477 *
roadways                               1.043e-06  3.361e-07   3.103 0.002305 **
refined_petrol_consumption            -5.306e-07  2.073e-07  -2.559 0.011516 *
birth_rate                             5.371e-02  3.002e-02   1.789 0.075709 .
airports                               3.951e-04  2.507e-04   1.576 0.117142
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 1.331 on 144 degrees of freedom
Multiple R-squared:  0.6737,  Adjusted R-squared:  0.6283
F-statistic: 14.86 on 20 and 144 DF,  p-value: < 2.2e-16
```

Algorithm 2: Result of the model

Visualize model 1.1 as Figure 3.

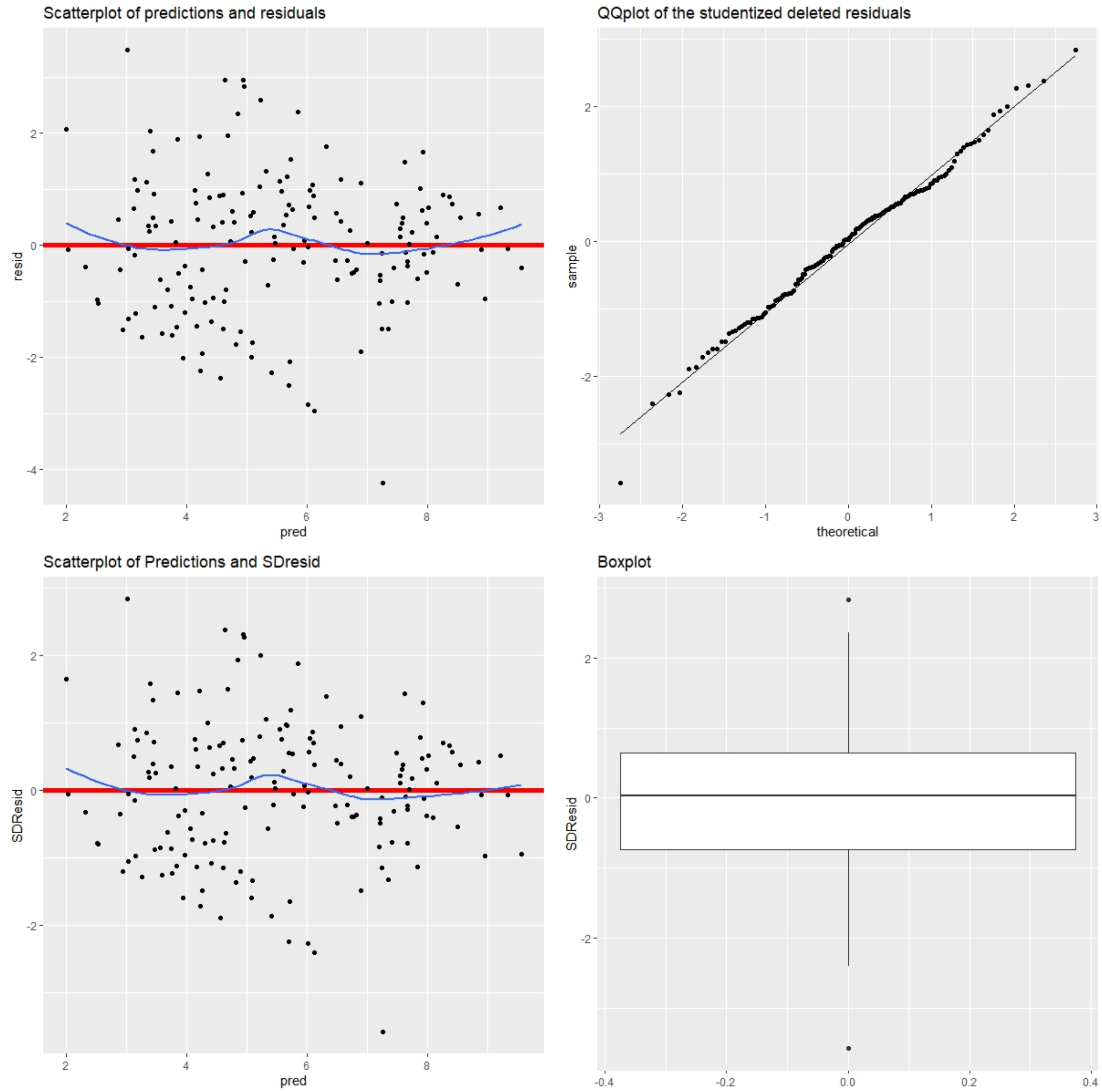

Figure 3: Properties of model 1.1

From the plots, we can conclude that there are outliers and distributions of errors have different variances.

Then, delete insignificant variables until all variables are significant. So, we drop "airports", and then mse and r squared do not change a lot.

Then, check multicolinearity of variables in the model. Roadways and refined petrol consumption are highly correlated and birth rate and infant mortality rate are highly correlated as well. To check significance of infant mortality rate, we need to delete birth rate because of their high correlation. The result of the model is as followed:

```
> summary(model1_3)

Call:
lm(formula = democracy_index ~ region + infant_mortality_rate +
    health_spend_pct_gdp + gdpPPP_percap + land_area + coastline +
    death_rate + roadways + refined_petrol_consumption, data = vars3)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4364 -0.8144  0.0241  0.7968  3.5577 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             4.363e+00  6.115e-01   7.135 4.19e-11 ***
regionAsia                              2.726e-01  9.789e-01   0.279 0.781006    
regionCentral America and the Caribbean 1.091e-01  5.072e-01   0.215 0.830041    
regionEurasia                          -1.806e+00  4.994e-01  -3.617 0.000410 ***
regionEurope                            1.131e-01  5.313e-01   0.213 0.831701    
regionMiddle East                      -2.004e+00  5.062e-01  -3.959 0.000117 ***
regionNorth America                     1.103e+00  1.367e+00   0.806 0.421287    
regionOceania                           1.757e+00  7.574e-01   2.320 0.021732 *  
regionSouth America                     1.094e+00  4.982e-01   2.197 0.029599 *  
regionSouth Asia                        4.627e-01  5.885e-01   0.786 0.432974    
regionSoutheast Asia                   -1.427e-01  4.996e-01  -0.286 0.775534    
infant_mortality_rate                  -4.059e-02  8.691e-03  -4.671 6.75e-06 ***
health_spend_pct_gdp                    1.488e-01  5.292e-02   2.812 0.005599 ** 
gdpPPP_percap                           2.446e-05  7.195e-06   3.400 0.000869 ***
land_area                              -3.033e-07  8.854e-08  -3.425 0.000797 ***
coastline                               1.682e-05  9.455e-06   1.779 0.077401 .  
death_rate                              1.126e-01  5.845e-02   1.926 0.056064 .  
roadways                                9.190e-07  3.327e-07   2.762 0.006483 ** 
refined_petrol_consumption             -3.359e-07  1.665e-07  -2.018 0.045461 *  
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 1.349 on 146 degrees of freedom
Multiple R-squared:  0.6605,	Adjusted R-squared:  0.6187 
F-statistic: 15.78 on 18 and 146 DF,  p-value: < 2.2e-16
```

Algorithm 3: Result of the model

Research if there is an interaction between roadways and refined petrol consumption. We will try 1) delete roadways, 2) refined petrol consumption, 3) both of them and 4) add an interaction. The result shows that performances of these models are almost same and we choose to drop both of them.

The result of t-test is as followed:

```
> summary(model1_7)

Call:
lm(formula = democracy_index ~ region + health_spend_pct_gdp + 
    gdpPPP_percap + land_area + coastline + death_rate + infant_mortality_rate, 
    data = vars3)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4318 -0.8706  0.0427  0.8245  3.5394 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             4.417e+00  6.196e-01   7.128 4.16e-11 ***
regionAsia                             -2.842e-01  9.086e-01  -0.313 0.754878    
regionCentral America and the Caribbean 1.103e-01  5.170e-01   0.213 0.831330    
regionEurasia                          -1.801e+00  5.088e-01  -3.539 0.000537 ***
regionEurope                            2.018e-01  5.407e-01   0.373 0.709457    
regionMiddle East                      -2.114e+00  5.130e-01  -4.120 6.29e-05 ***
regionNorth America                     3.684e-01  1.125e+00   0.327 0.743755    
regionOceania                           1.737e+00  7.720e-01   2.250 0.025932 *  
regionSouth America                     1.081e+00  5.072e-01   2.131 0.034749 *  
regionSouth Asia                        1.042e+00  5.605e-01   1.859 0.065011 .  
regionSoutheast Asia                   -1.932e-01  5.001e-01  -0.386 0.699852    
health_spend_pct_gdp                    1.438e-01  5.189e-02   2.772 0.006284 ** 
gdpPPP_percap                           2.227e-05  7.254e-06   3.070 0.002545 ** 
land_area                              -2.341e-07  7.665e-08  -3.053 0.002683 ** 
```

```
coastline                                 1.900e-05  8.469e-06   2.243 0.026369 *
death_rate                                1.139e-01  5.946e-02   1.916 0.057354 .
infant_mortality_rate                    -4.129e-02  8.858e-03  -4.662 6.94e-06 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 1.375 on 148 degrees of freedom
Multiple R-squared:  0.6422,  Adjusted R-squared:  0.6036
F-statistic: 16.61 on 16 and 148 DF,  p-value: < 2.2e-16
```

Algorithm 4: Result of the model

To solve heteroscedasticity, we try weighted least squares. The result is shown as follows:

```
summary(model1_8)

Call:
lm(formula = democracy_index ~ region + health_spend_pct_gdp +
    gdpPPP_percap + land_area + coastline + death_rate + infant_mortality_rate,
    data = vars3, weights = cal.weights)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.1149 -0.7665  0.0369  0.8359  2.9334

Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                               4.435e+00  6.343e-01   6.992 8.66e-11 ***
regionAsia                               -1.790e-01  7.849e-01  -0.228 0.819917
regionCentral America and the Caribbean  -7.394e-02  5.076e-01  -0.146 0.884375
regionEurasia                            -1.699e+00  5.683e-01  -2.990 0.003272 **
regionEurope                              3.435e-01  5.107e-01   0.673 0.502226
regionMiddle East                        -2.395e+00  5.702e-01  -4.200 4.59e-05 ***
regionNorth America                       5.687e-02  1.004e+00   0.057 0.954903
regionOceania                             1.724e+00  6.202e-01   2.780 0.006142 **
regionSouth America                       1.056e+00  4.727e-01   2.234 0.026960 *
regionSouth Asia                          9.701e-01  5.737e-01   1.691 0.092975 .
regionSoutheast Asia                     -3.222e-01  5.131e-01  -0.628 0.531052
health_spend_pct_gdp                      1.564e-01  4.783e-02   3.270 0.001339 **
gdpPPP_percap                             2.822e-05  6.447e-06   4.377 2.26e-05 ***
land_area                                -1.991e-07  7.138e-08  -2.789 0.005976 **
coastline                                 1.614e-05  5.553e-06   2.908 0.004203 **
death_rate                                6.804e-02  5.509e-02   1.235 0.218762
infant_mortality_rate                    -3.546e-02  9.231e-03  -3.841 0.000181 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 1.312 on 148 degrees of freedom
Multiple R-squared:  0.6983,  Adjusted R-squared:  0.6657
F-statistic: 21.41 on 16 and 148 DF,  p-value: < 2.2e-16
```

Algorithm 5: Result of the model

r squared becomes higher, but death rate becomes totally unsiginificant. So delete the variable.

Then, try three measures to drop outliers. But the outliers are not significant, so the model does not change.

2 PART II

2.1 Introduction

The goal of the section is to predict the life expectancy. In the regression model, life expectancy is chosen as the response variable. We will use multiple linear regression at first, tring to find the best-in-class model by selecting appropriate variables. In order to improve the performance of the prediction, a few different approaches will be employed.

2.2 Statistical Model

After analysis and trials, the model and its results are as followed:

```
> summary(model2_3)

Call:
lm(formula = life_exp_at_birth ~ infant_mortality_rate + death_rate +
    region + urbanization + birth_rate + health_spend_pct_gdp,
    data = training_set2_3)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6996 -0.9113  0.0546  1.2345  6.2050

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           81.03005    1.42942  56.687  < 2e-16 ***
infant_mortality_rate -0.16273    0.02262  -7.195 5.23e-11 ***
death_rate            -0.82084    0.09034  -9.086 2.03e-15 ***
regionEurope           3.55124    0.64153   5.536 1.76e-07 ***
regionSouth Asia       2.53955    0.97126   2.615 0.010038 *
urbanization           0.05553    0.01161   4.784 4.79e-06 ***
birth_rate            -0.23944    0.04720  -5.073 1.39e-06 ***
health_spend_pct_gdp   0.31302    0.08158   3.837 0.000197 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 2.18 on 124 degrees of freedom
Multiple R-squared:  0.9318,  Adjusted R-squared:  0.928
F-statistic: 242.2 on 7 and 124 DF,  p-value: < 2.2e-16

> anova(model2_3)
Analysis of Variance Table

Response: life_exp_at_birth
                       Df Sum Sq Mean Sq  F value    Pr(>F)
infant_mortality_rate   1 7262.8  7262.8 1528.064 < 2.2e-16 ***
death_rate              1  181.2   181.2   38.120 8.716e-09 ***
region                  2  246.7   123.3   25.948 3.859e-10 ***
urbanization            1  189.4   189.4   39.847 4.449e-09 ***
birth_rate              1  106.5   106.5   22.402 5.927e-06 ***
health_spend_pct_gdp    1   70.0    70.0   14.724 0.0001972 ***
Residuals             124  589.4     4.8
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
```

Algorithm 6: Result of the model

Variables are infant mortality rate, death rate, region, urbanization, birth rate, health spend pct gdp. Multiple R-squared is 0.9318 and Adjusted R-squared is 0.928, which means the performance of the model is very good. The AIC of the model is -590, and the MSPE is 3.71. Further diagnostics on multicolinearity, Heteroscedasticity, log transformation, and outliers will be discussed in the next section.

2.3 Appendix

2.3.1 *Model (including Diagnostics)*

In this problem, firstly split the data set into 80% training set and 20% test set. Then, choose variables using forward selection with AIC. The covariates are infant mortality rate, gdpPPP percap, death rate, region, urbanization, birth rate, health spend pct gdp, continent, land area, coastline.

```
> summary(model2_1)

Call:
lm(formula = life_exp_at_birth ~ infant_mortality_rate + gdpPPP_percap +
    death_rate + region + urbanization + birth_rate + health_spend_pct_gdp +
    continent + land_area + coastline, data = training_set2_1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3569 -1.1371  0.0324  1.1719  5.5983

Coefficients: (4 not defined because of singularities)
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              7.869e+01  1.598e+00  49.241  < 2e-16 ***
infant_mortality_rate                   -1.506e-01  2.083e-02  -7.228 6.40e-11 ***
gdpPPP_percap                            3.626e-06  1.327e-05   0.273 0.785155
death_rate                              -9.866e-01  1.033e-01  -9.550 3.62e-16 ***
regionAsia                               8.535e+00  1.926e+00   4.431 2.19e-05 ***
regionCentral America and the Caribbean  1.371e+00  9.198e-01   1.491 0.138788
regionEurasia                            3.354e+00  9.860e-01   3.401 0.000930 ***
regionEurope                             6.383e+00  8.956e-01   7.127 1.06e-10 ***
regionMiddle East                        1.458e+00  1.300e+00   1.122 0.264402
regionNorth America                      5.701e+00  2.550e+00   2.235 0.027378 *
regionOceania                            5.007e+00  1.312e+00   3.816 0.000223 ***
regionSouth America                      1.374e+00  8.475e-01   1.622 0.107662
regionSouth Asia                         6.183e+00  1.488e+00   4.155 6.39e-05 ***
regionSoutheast Asia                     4.762e+00  1.422e+00   3.349 0.001105 **
urbanization                             6.424e-02  1.256e-02   5.112 1.32e-06 ***
birth_rate                              -1.456e-01  4.801e-02  -3.032 0.003018 **
health_spend_pct_gdp                     2.427e-01  8.349e-02   2.907 0.004403 **
continentAsia                           -2.343e+00  1.155e+00  -2.028 0.044965 *
continentEurope                                 NA         NA      NA       NA
continentNorth America                          NA         NA      NA       NA
continentOceania                                NA         NA      NA       NA
continentSouth America                          NA         NA      NA       NA
land_area                               -3.492e-07  1.325e-07  -2.635 0.009606 **
coastline                                1.645e-05  1.356e-05   1.213 0.227708
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 1.95 on 112 degrees of freedom
Multiple R-squared:  0.9508,  Adjusted R-squared:  0.9424
F-statistic: 113.8 on 19 and 112 DF,  p-value: < 2.2e-16

> anova(model2_1)
Analysis of Variance Table

Response: life_exp_at_birth
                      Df Sum Sq Mean Sq   F value    Pr(>F)
infant_mortality_rate  1 7262.8  7262.8 1910.5730 < 2.2e-16 ***
gdpPPP_percap          1  219.7   219.7   57.8041 9.602e-12 ***
death_rate             1  144.9   144.9   38.1206 1.094e-08 ***
region                10  364.8    36.5    9.5975 2.131e-11 ***
urbanization           1  117.1   117.1   30.8078 1.935e-07 ***
birth_rate             1   34.6    34.6    9.1076  0.003152 **
```

```
health_spend_pct_gdp    1   39.4    39.4   10.3749  0.001672 **
continent               1    7.2     7.2    1.8936  0.171542
land_area               1   23.8    23.8    6.2740  0.013691 *
coastline               1    5.6     5.6    1.4712  0.227708
Residuals             112  425.8     3.8
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1
```

Algorithm 7: Result of the model

However, this primary version of model has some critical shortcomes. There are too many variables, which might cause overfitting. Besides, many of the variables are not so significant and covariate. By sampling training set randomly several times, we can get following conclusions:

By F test, we find gdpPPP percap, continent, land area and coastline are not so significant and should be dropped therefore.

By t test, we find some regions are not significant and can be supposed as "other regions" therefore. Also, it should be better to set "other regions" as baseline.

The visulization of the model in training set is shown as below: The result in the training set is not bad, especially when y is large. Comparing y hat and residuals results that there's no normality. QQ plot shows that the distribution has a short tail. When we test the model in the test set, the MSE in the test set is 13.26. Because the data size in the test set is very small, MSE will change a lot because of the influence of degree of freedom, so it would be better to use the biased MSE without considering the degree of freedom. The biased MSE of this model in the test set is 5.63, which is still larger than the training set. Our model overfits the training set.

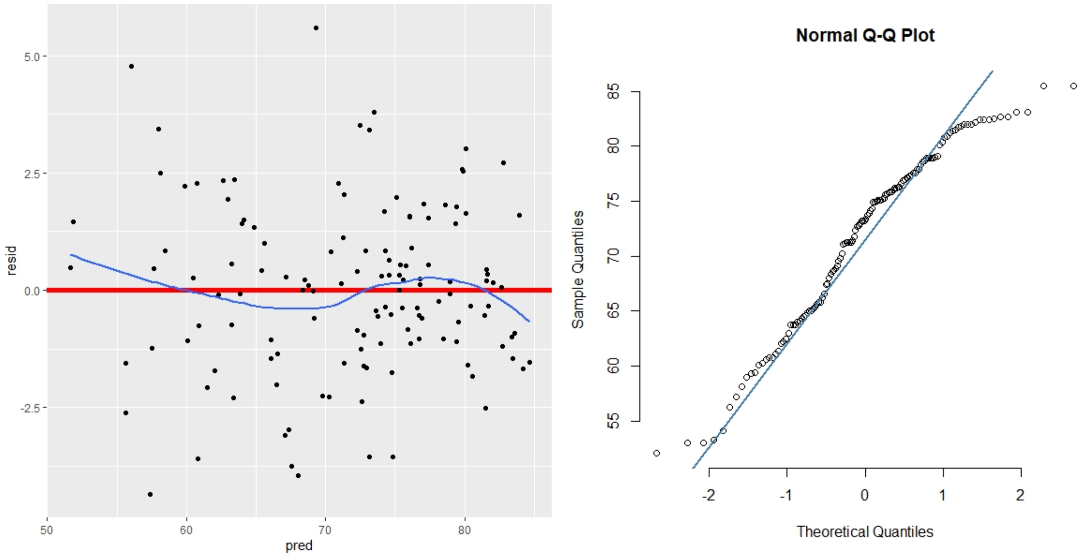

Figure 4: Properties of model 2.1

The visulization of the prediction is shown in Figure 5:

Then we try to modify the mlr model by deleting variables. We find some regions are not significant and we will classify some regions as "Others" therefore. We continue to select variables until every variable in the model is significant in training set.

```

> summary(model2_3)

Call:
lm(formula = life_exp_at_birth ~ infant_mortality_rate + death_rate +
```

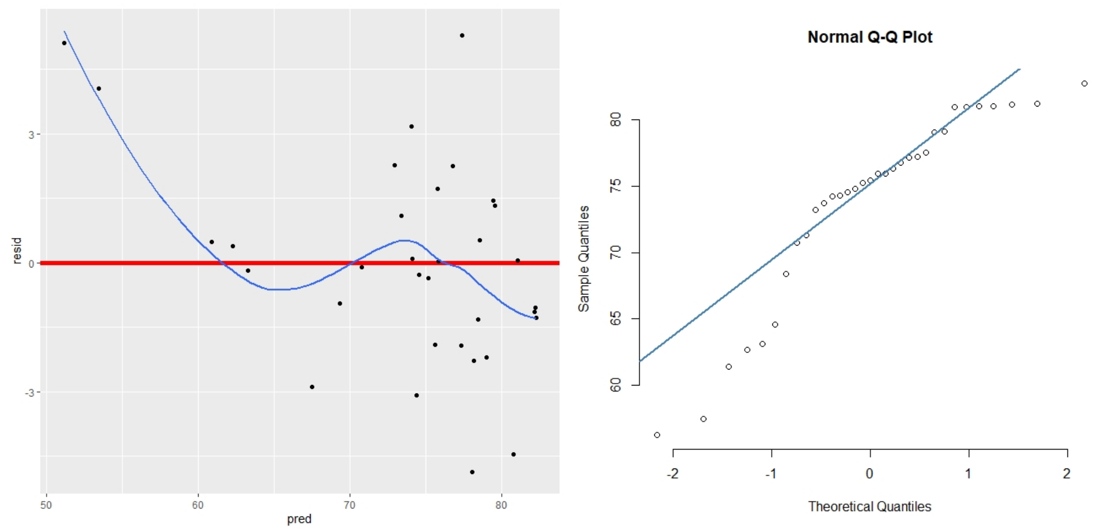

Figure 5: Prediction of model 2.1

```
    region + urbanization + birth_rate + health_spend_pct_gdp,
    data = training_set2_3)

Residuals:
    Min      1Q   Median      3Q     Max
-6.6996 -0.9113  0.0546  1.2345  6.2050

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           81.03005    1.42942  56.687  < 2e-16 ***
infant_mortality_rate -0.16273    0.02262  -7.195 5.23e-11 ***
death_rate            -0.82084    0.09034  -9.086 2.03e-15 ***
regionEurope           3.55124    0.64153   5.536 1.76e-07 ***
regionSouth Asia       2.53955    0.97126   2.615 0.010038 *
urbanization           0.05553    0.01161   4.784 4.79e-06 ***
birth_rate            -0.23944    0.04720  -5.073 1.39e-06 ***
health_spend_pct_gdp   0.31302    0.08158   3.837 0.000197 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 2.18 on 124 degrees of freedom
Multiple R-squared:  0.9318,  Adjusted R-squared:  0.928
F-statistic: 242.2 on 7 and 124 DF,  p-value: < 2.2e-16

> anova(model2_3)
Analysis of Variance Table

Response: life_exp_at_birth
                       Df Sum Sq Mean Sq  F value    Pr(>F)
infant_mortality_rate   1 7262.8  7262.8 1528.064 < 2.2e-16 ***
death_rate              1  181.2   181.2   38.120 8.716e-09 ***
region                  2  246.7   123.3   25.948 3.859e-10 ***
urbanization            1  189.4   189.4   39.847 4.449e-09 ***
birth_rate              1  106.5   106.5   22.402 5.927e-06 ***
health_spend_pct_gdp    1   70.0    70.0   14.724 0.0001972 ***
Residuals             124  589.4     4.8
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1    1
```

Algorithm 8: Result of the model

We test the new model in the test set. This time MSE is 4.71, and the biased MSE is 3.71. Both of them are much smaller than the previous model.

We further try to diagnose the multicolinearity. The covariance matrix is shown as below, which shows that infant mortality rate and birth rate are highly correlated. Compare the three models: with both two, without birth rate and without infant mortality rate. The performance of the first model is the best, so we cannot delete either of the variables. We then do diagnostics on Heteroscedasticity. Weighted Least Squares is employed. However, the new model performs a little bit worse than unweighted one. The sample size of test set is too small to conclude which model is better. But variance of residuals is smaller in weighted model. Theoratically, the model with weight can perform better when sample size is large. The reason of bad performance of weighted model might be the exsitence of outliers. We also tried logY as the response variable. It performs well but not so well in test set. Finally, we tried to diagnose the ourliers. A robust regression is employed to deal with this issue.

```

> summary(model2_9)

Call: rlm(formula = life_exp_at_birth ~ infant_mortality_rate + death_rate +
    region + urbanization + birth_rate + health_spend_pct_gdp,
    data = training_set2_3)
Residuals:
      Min         1Q    Median        3Q       Max
-7.193251  -0.977662  0.008508  1.020175  6.155282

Coefficients:
                      Value    Std. Error  t value
(Intercept)            82.1256   1.2127     67.7213
infant_mortality_rate  -0.1553   0.0192     -8.0946
death_rate             -0.8156   0.0766    -10.6412
regionEurope            3.3948   0.5443      6.2375
regionSouth Asia        2.0879   0.8240      2.5339
urbanization            0.0440   0.0098      4.4662
birth_rate             -0.2738   0.0400     -6.8384
health_spend_pct_gdp    0.3317   0.0692      4.7930

Residual standard error: 1.486 on 124 degrees of freedom
> anova(model2_9)
Analysis of Variance Table

Response: life_exp_at_birth
                      Df Sum Sq Mean Sq F value Pr(>F)
infant_mortality_rate  1 6298.6  6298.6
death_rate             1  117.9   117.9
region                 2  228.5   114.2
urbanization           1  122.9   122.9
birth_rate             1  124.8   124.8
health_spend_pct_gdp   1   72.8    72.8
Residuals                 596.7
```

Algorithm 9: Result of the model

In the test set, MSE this time is 4.83, and the biased MSE is 3.81. The result is almost the same as Model 2.3. The visulization is shown as below.

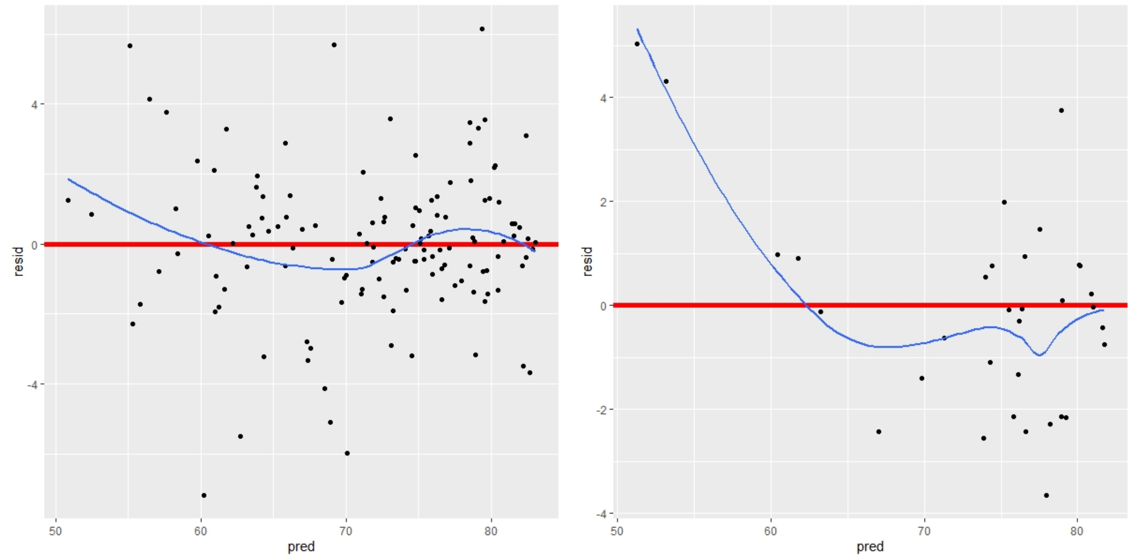

Figure 6: Properties of model 2.9