Analysis of Global Life Expectancy

Dingkun Yang, Echo Chen, Andrew Kroening, Pooja Kabber

November 22nd, 2022

Abstract

A few sentences describing the purpose of the analysis, the data, and key results.

The analysis conducted below (is based on) understanding which factors influence life expectancy and the development status of countries around the world. The dataset used for this research consisting of national disease, economic, and social (factors) has been collected by the World Health Organization (WHO). The WHO compiles data on thousands of variables for as many countries as feasible and presents them for analysis.

(key results)

Introduction

Provide more background on the data and research questions. Be sure to cite the data and background information appropriately (APA style is fine)

This particular dataset contains national-level observations of variables related to life expectancy around the globe for a period spanning the early portion of the 21st century. The complete dataset includes observations beginning in the year 2000 and ending in the year 2015. As a full dataset, there are 2,938 observations for 22 variables. Practically, each country has approximately one observation each year, averaging 183 for each of the 16 years encompassed by the data. The dataset effectively contains 20 variables for each country and year combination, covering significant disease, economic, and social factors.

Research Questions

Below are the questions we aim to answer based on this analysis:

Question #1 (Prediction)

"How did major disease, economic, and social factors impact life expectancy around the globe in 2014?"

Question #2 (Inference)

"How did disease and mortality rates, along with national economic factors, contribute to a country's development status in 2014?"

(cite data and background info)

Methods

Describe the process you used to conduct analysis. This includes EDA and any relevant data cleaning information (e.g., did you exclude missing values? If so, how many? Did you collapse categories for any variables?) Then describe the models you fit, and any changes you made to improve model fit (e.g., did you exclude any influential points? Did you do have to address multicollinearity issues? Did you transform

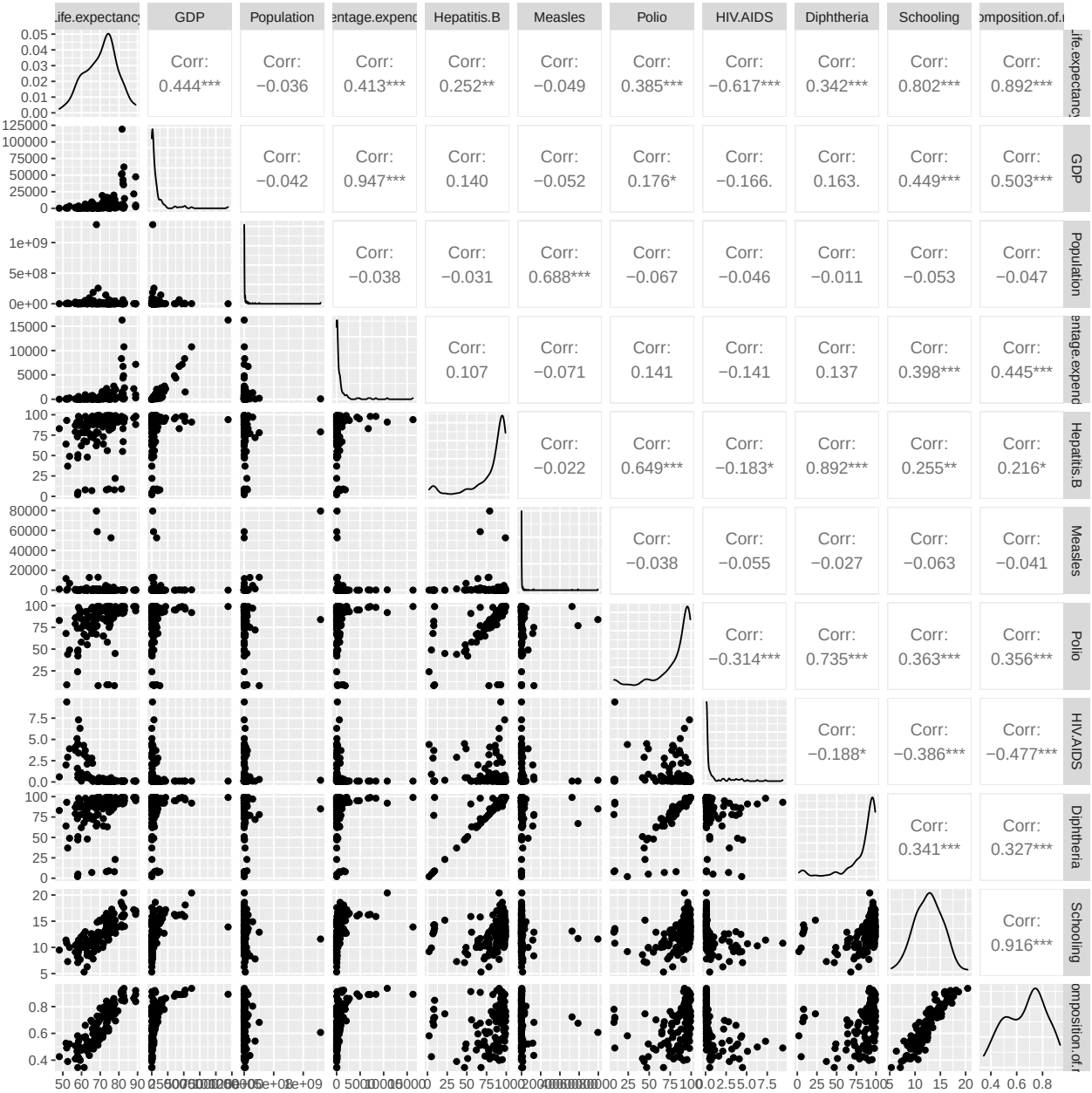
any variables?). Also describe model diagnostics. The organization of this section may depend on your particular dataset/analysis, but you may want to break it into subsections such as "Data," "Models," and "Model assessment." Note that you do not present any results in this section.

Data

1. Missing data - complete case and last two methods
2. Other data cleaning

Subsetted data for only 2014

3. EDA plots from part 1



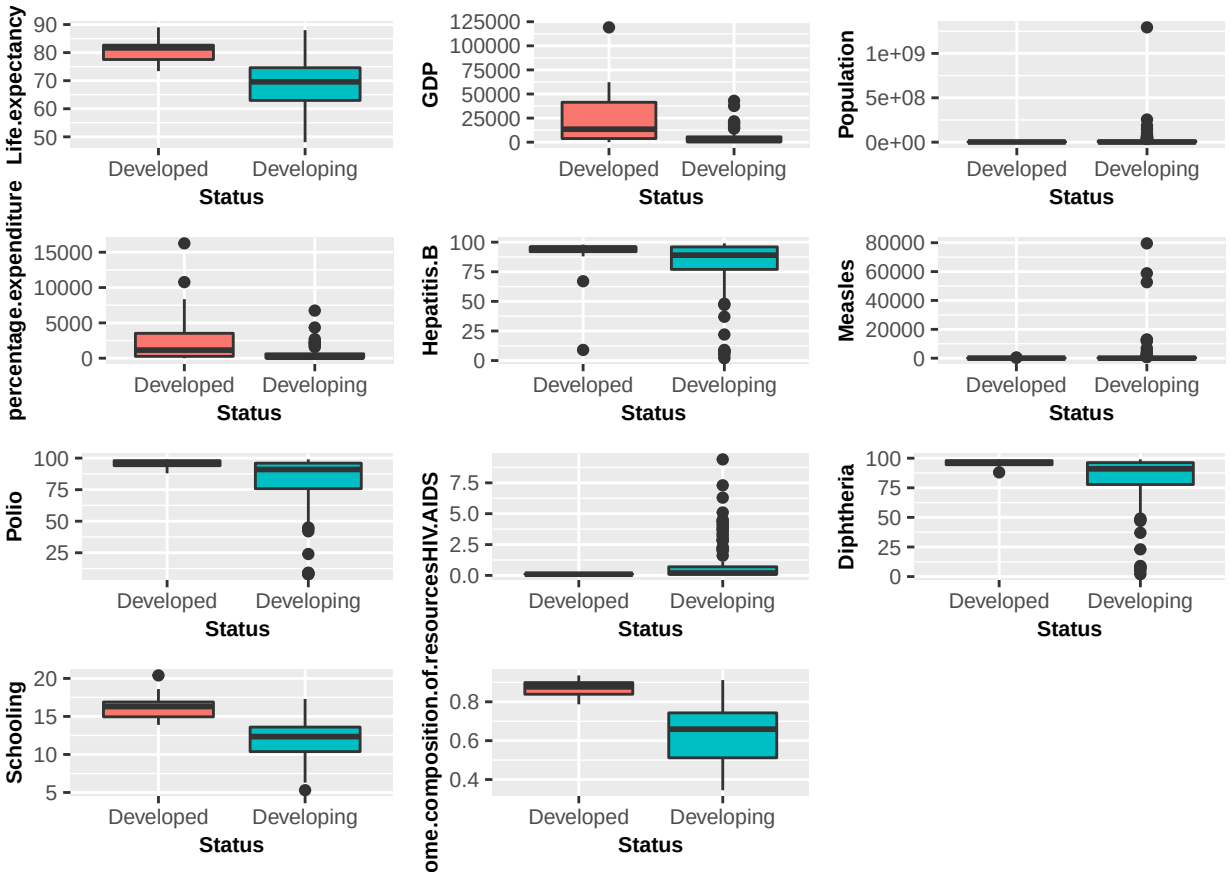


Table 1: Summary of Variables

Statistic	N	Mean	St. Dev.	Min	Max
Population	131	22,269,096.000	116,699,866.000	41.000	1,293,859,294.000
Life.expectancy	131	70.520	8.605	48.100	89.000
percentage.expenditure	131	850.874	2,071.444	0.443	16,255.160
Measles	131	2,042.863	9,842.341	0	79,563
Polio	131	83.496	20.966	8	99
HIV.AIDS	131	0.810	1.562	0.100	9.400
GDP	131	7,256.847	14,741.400	12.277	119,172.700
Schooling	131	12.676	2.750	5.300	20.400
Income.composition.of.resources	131	0.670	0.151	0.345	0.936
BMI	131	40.476	20.734	2.000	77.100
Total.expenditure	131	6.107	2.533	1.210	13.730

Models

4. Model description

The model used is linear regression.

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 3 in
## model.matrix: no columns are assigned
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Population + Life.expectancy +
##     percentage.expenditure + Measles + Polio + HIV.AIDS + GDP +
##     Schooling + Income.composition.of.resources + BMI + Total.expenditure,
##     data = df_life_expectancy_2014)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.910  -1.885   0.186   1.848   8.598
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.262e+01  2.751e+00  15.492  < 2e-16 ***
## StatusDeveloping                 -9.977e-01  1.070e+00  -0.932   0.3531
## Population                        1.394e-09  3.545e-09   0.393   0.6949
## percentage.expenditure            5.012e-04  4.793e-04   1.046   0.2978
## Measles                          -2.977e-05  4.247e-05  -0.701   0.4846
## Polio                             1.008e-02  1.605e-02   0.628   0.5310
## HIV.AIDS                         -1.365e+00  2.267e-01  -6.020 1.99e-08 ***
## GDP                              -6.893e-05  6.930e-05  -0.995   0.3219
## Schooling                        -2.175e-01  2.811e-01  -0.774   0.4407
## Income.composition.of.resources   4.547e+01  5.906e+00   7.699 4.47e-12 ***
## BMI                              -5.827e-03  1.941e-02  -0.300   0.7645
## Total.expenditure                 2.722e-01  1.329e-01   2.047   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.404 on 119 degrees of freedom
## Multiple R-squared:  0.8567, Adjusted R-squared:  0.8435
## F-statistic: 64.69 on 11 and 119 DF,  p-value: < 2.2e-16
```

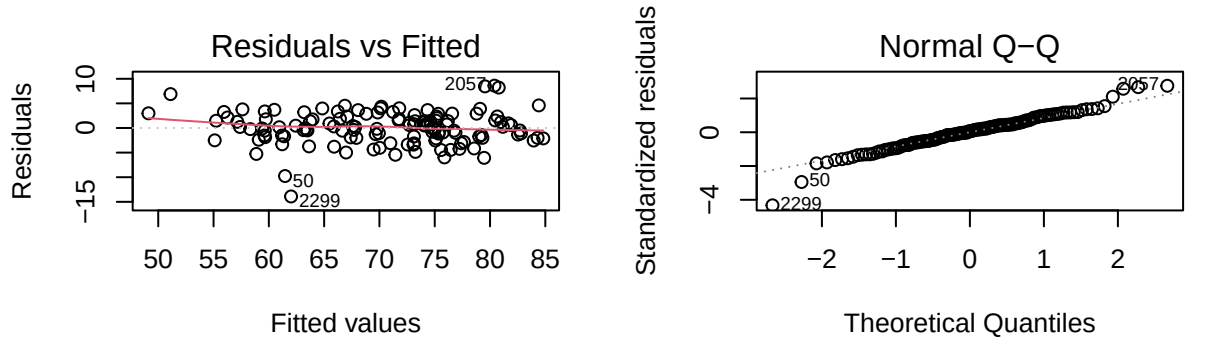
```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = subset(df_life_expectancy_cc_2014,
##     select = -c(Country, Year)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4098  -1.7264  -0.0392   1.7715   8.3880
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.122e+01  3.314e+00  15.457  < 2e-16 ***
## StatusDeveloping                 -1.170e+00  1.035e+00  -1.130 0.261006
## Adult.Mortality                  -1.724e-02  4.148e-03  -4.157 6.36e-05 ***
## infant.deaths                     8.287e-02  5.619e-02   1.475 0.143057
## Alcohol                           5.674e-03  9.749e-02   0.058 0.953689
## percentage.expenditure            4.627e-04  4.639e-04   0.997 0.320716
## Hepatitis.B                       1.205e-02  2.808e-02   0.429 0.668582
## Measles                          -3.361e-05  4.823e-05  -0.697 0.487345
## BMI                              -7.576e-03  2.000e-02  -0.379 0.705531
## under.five.deaths                -6.014e-02  3.838e-02  -1.567 0.119989
## Polio                            -8.746e-03  2.117e-02  -0.413 0.680327
```

```
## Total.expenditure                2.878e-01  1.274e-01   2.259 0.025833 *
## Diphtheria                       7.644e-03  3.445e-02   0.222 0.824805
## HIV.AIDS                        -8.363e-01  2.470e-01  -3.385 0.000984 ***
## GDP                             -5.980e-05  6.656e-05  -0.898 0.370911
## Population                      -1.729e-09  6.804e-09  -0.254 0.799816
## thinness..1.19.years            -1.300e-01  2.267e-01  -0.574 0.567462
## thinness.5.9.years               5.458e-03  2.227e-01   0.025 0.980489
## Income.composition.of.resources  3.597e+01  6.228e+00   5.775 7.11e-08 ***
## Schooling                       -1.617e-01  2.740e-01  -0.590 0.556279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.186 on 111 degrees of freedom
## Multiple R-squared:  0.8829, Adjusted R-squared:  0.8629
## F-statistic: 44.06 on 19 and 111 DF,  p-value: < 2.2e-16
```

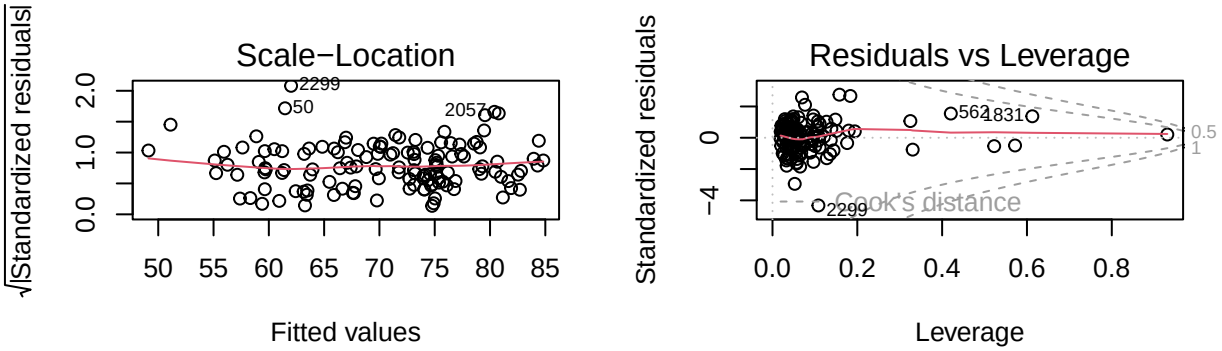
```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     under.five.deaths + Total.expenditure + HIV.AIDS + Income.composition.of.resources,
##     data = subset(df_life_expectancy_cc_2014, select = -c(Country,
##         Year)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.568  -1.569  -0.127   1.561  10.060
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     47.860575   2.066580  23.159  < 2e-16 ***
## Adult.Mortality                 -0.017405   0.003866  -4.502 1.53e-05 ***
## infant.deaths                    0.042287   0.029527   1.432 0.154625
## under.five.deaths               -0.033561   0.022614  -1.484 0.140331
## Total.expenditure                0.349095   0.111930   3.119 0.002258 **
## HIV.AIDS                        -0.809261   0.231621  -3.494 0.000661 ***
## Income.composition.of.resources 35.911609   2.502726  14.349  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.096 on 124 degrees of freedom
## Multiple R-squared:  0.8765, Adjusted R-squared:  0.8705
## F-statistic: 146.7 on 6 and 124 DF,  p-value: < 2.2e-16
```

```
## Warning in model.matrix.default(object, data = structure(list(Life.expectancy =
## c(59.9, : the response appeared on the right-hand side and was dropped
```

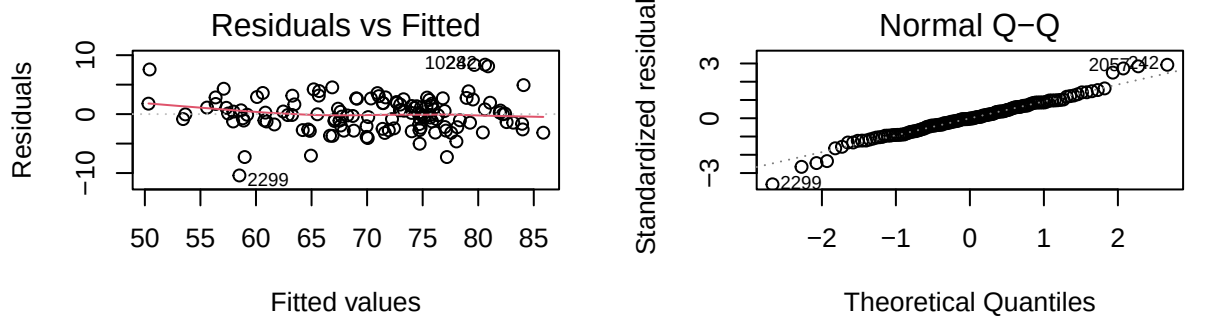
```
## Warning in model.matrix.default(object, data = structure(list(Life.expectancy =
## c(59.9, : problem with term 3 in model.matrix: no columns are assigned
```


Residuals vs Fitted
Residuals
Fitted values
2057
50
2299

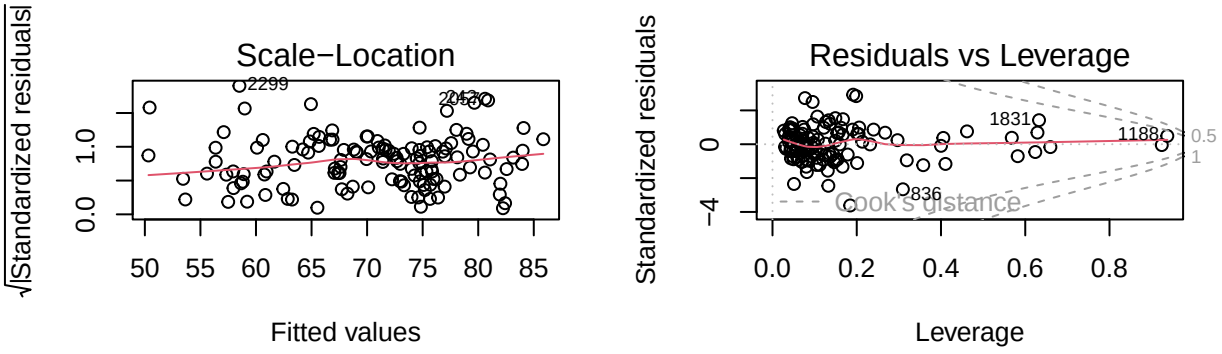
Normal Q−Q
Standardized residuals
Theoretical Quantiles
2057
50
2299


Scale−Location
√|Standardized residuals|
Fitted values
2299
50
2057

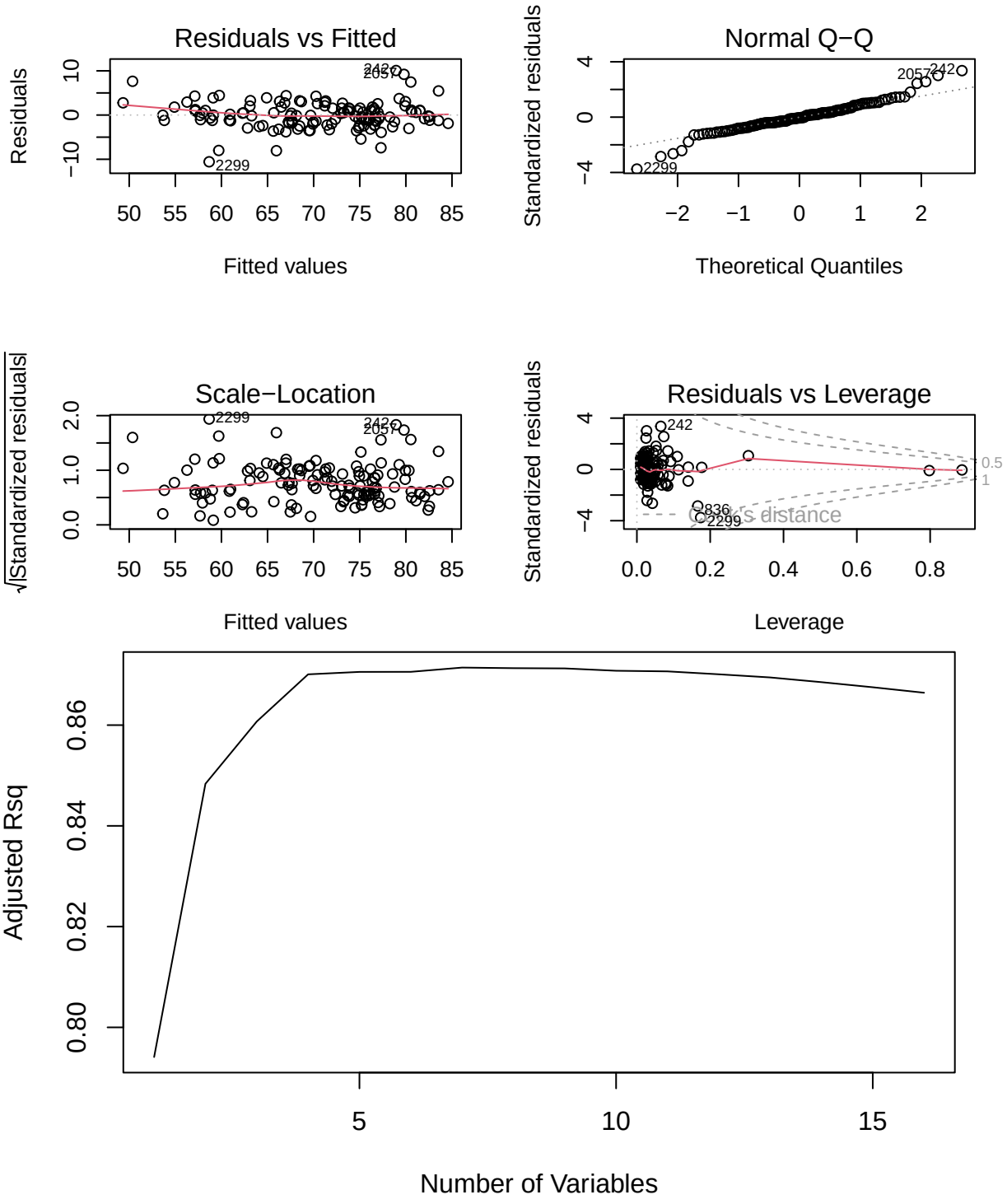
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
562
1831
2299


Residuals vs Fitted
Residuals
Fitted values
2299

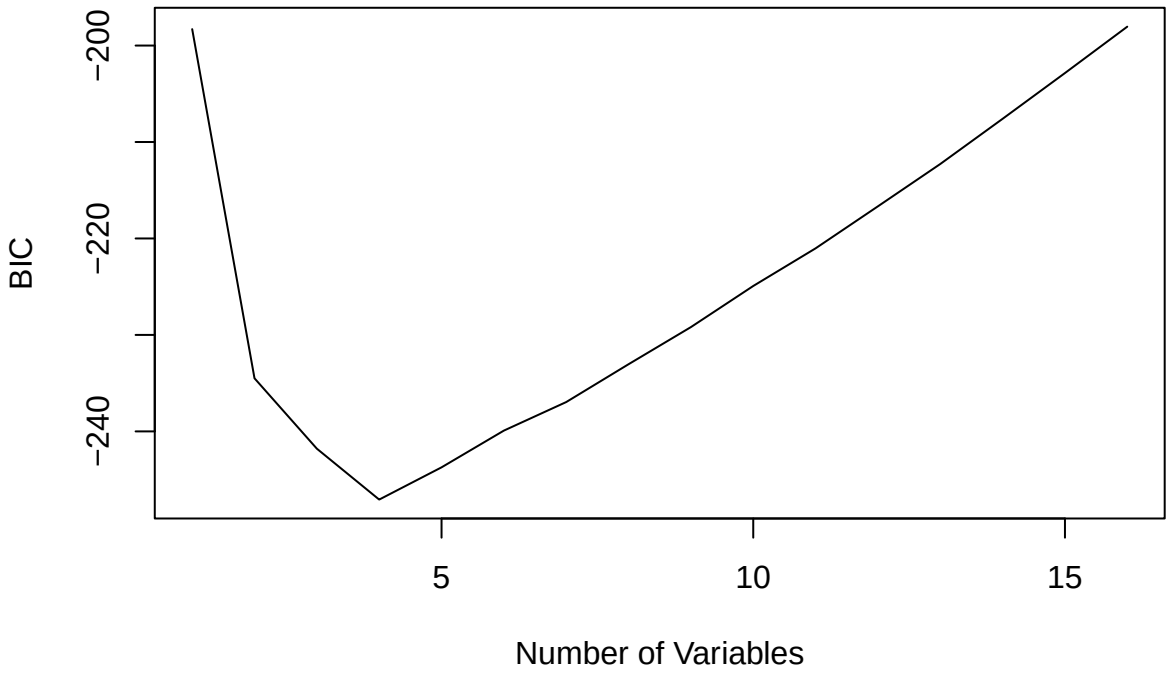
Normal Q−Q
Standardized residuals
Theoretical Quantiles
2299


Scale−Location
√|Standardized residuals|
Fitted values
2299

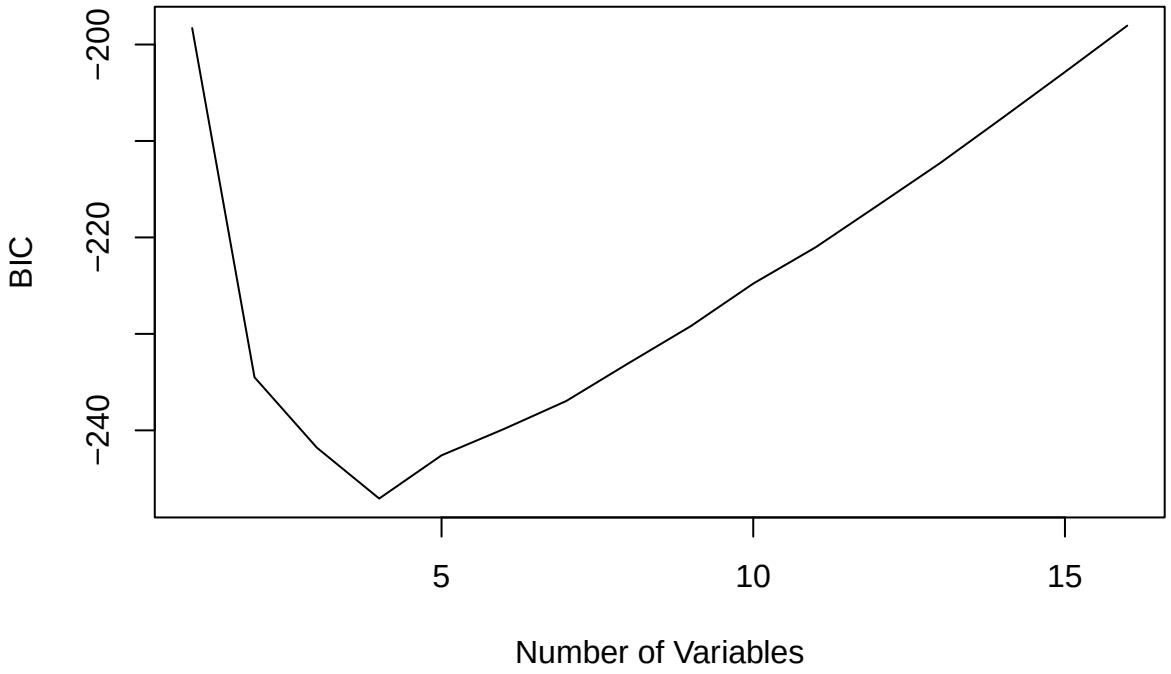
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1831
836


Residuals vs Fitted
Residuals
Fitted values
2299
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
242
836
Adjusted Rsq
Number of Variables

```
## [1] 7
```




```
## [1] 4
```




```
## [1] 4
```

Model Assessment

5. Model assessment (transformations, influential points, multicollinearity), model evaluation

Results

Here you should present results for all aspects of the analysis. The structure of this section should mirror the structure of the methods section. For example, you can start with a few key EDA results (e.g., a table of descriptive statistics), then present model results, then address assessment. This is the section where you will primarily refer to tables and figures. You should have at least 1 figure for each research question that illustrates a key result of the analysis.

Data

1. EDA

Models

2. Model results- Model summary table, confidence intervals, interpretation. (Why are we primarily referring to tables and figures here?)
3. Visualisations for key results

Model Assessment

(?)

Conclusion

Describe the key takeaways from your analysis, limitations, and future work that can be done to advance knowledge in this area.