Factors affecting the GNI per capita—birth rates, death rates, infant mortality rates, life expectancies, and regions

Ophelia Li

Math 7360 - Data Analysis

1. Introduction and Data Selection

The Gross National Income (GNI) per capita is the dollar value of a country's final income in a year, divided by its population. The GNI per capita is a good indicator of a country's economic strengths and the standard of living of the average citizen. Several factors are often mentioned to influence GNI per capita. In this report, we want to investigate how birth rates, death rates, life expectancies, infant mortality rates, and regions affect the GNI per capita for 173 countries.

The data was taken from the World Bank Open Data (https://data.worldbank.org/) and was from the year 2018. The data consist of 8 variables for 173 countries.

- birthrate: live birth rate per 1,000 of population
- deathrate: death rate per 1,000 of population
- lifeexpM: life expectancy at birth for male in years
- lifeexpF: life expectancy at birth for male in years
- infantdeaths: mortality rate of infants per 1,000 live births
- GNIpercapita: gross national income per capita in US dollars
- region: each country is categorized into a geographic region as follows:
  1. East Asia & Pacific
  2. South Asia
  3. Europe & Central Asia
  4. North America
  5. Latin America
  6. Middle East & North Africa
  7. Sub-Saharan Africa
- country: country names

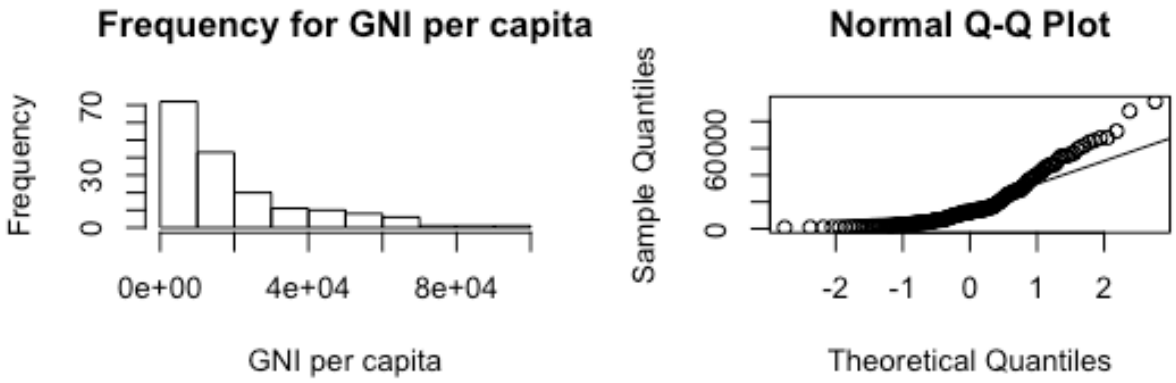
2. Objective

2.1 Investigate if GNI per capita can be well predicted by the factors mentioned in the dataset—birth rate, death rate, infant death rate, life expectancy, and region.

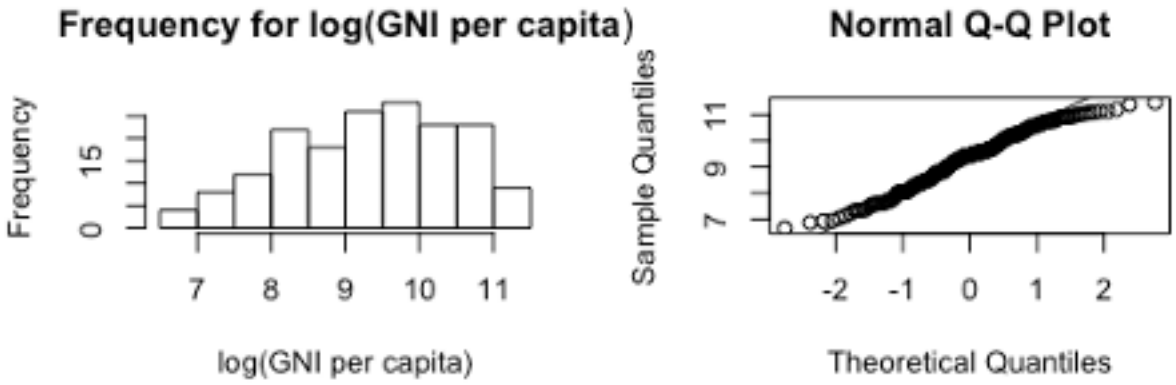
2.2 Apply transformations to improve the predictability of the model.

3. Summary of the Data

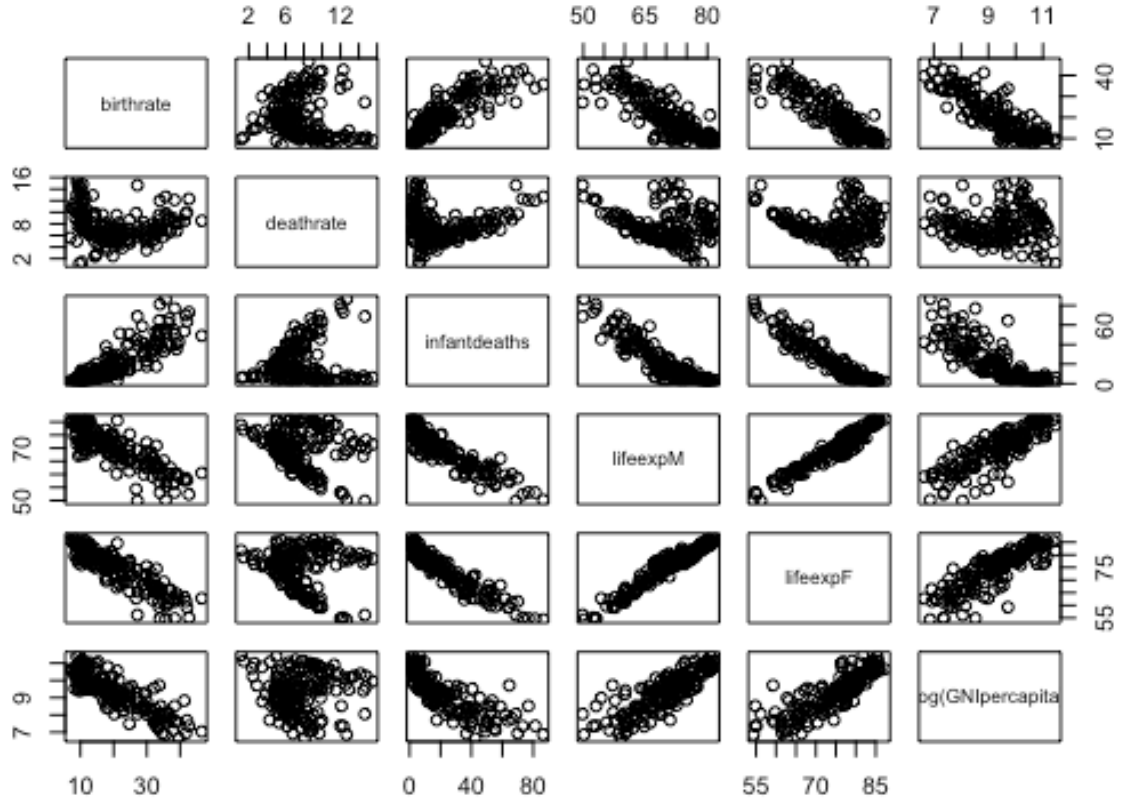
We begin by looking at the distribution of the GNI per capita. The cumulative frequency graph of the GNI per capita shows a high right-skewness and the Q-Q plot shows that many points do not fall about a straight line.



A transformation in similar situations is to use the logarithm of the GNI per capita. The distribution of the GNI per capita becomes more centered and points fall about a straight line on the Q-Q plot. We will then use log(GNI per capita) as the response variable.



Now we examine covariates. Pairwise scatterplots of the predictors are shown below:



There is collinearity between "lifeexpM" and "lifeexpF." We can confirm it by calculating the correlation between predictors, as shown in the following table:

	birthrate	deathrate	lifeexpM	lifeexpF	infantdeaths
birthrate	1.000	-0.144	-0.836	-0.887	0.865
deathrate	-0.144	1.000	-0.206	-0.103	0.099
lifeexpM	-0.836	-0.206	1.000	0.966	-0.902
lifeexpF	-0.887	-0.103	0.966	1.000	-0.945
infantdeaths	0.865	0.099	-0.902	-0.945	1.000

Thus, we eliminate "lifeexpM" and "lifeexpF" variables and introduce two new variables that can better address the effect of life expectancy on the GNI per capita:

- life.exp.avg: the average life expectancy at birth for both males and females
- life.exp.diff: the difference in average life expectancy between males and females

4. Model Selections and Results

4.1 Linear Model

We start from a large model including all the predictor variables and then do a sequential backward search using the step function to find the best-fitted model. The best-fitted model selected is:

```
log(GNIpercapita) ~ birthrate + life.exp.avg + region
```

And an adjusted R-squared of 0.7979 indicates relatively high predictability. We will now look at ways to improve the adjusted R-squared.

4.2.1 Adding Interaction Terms

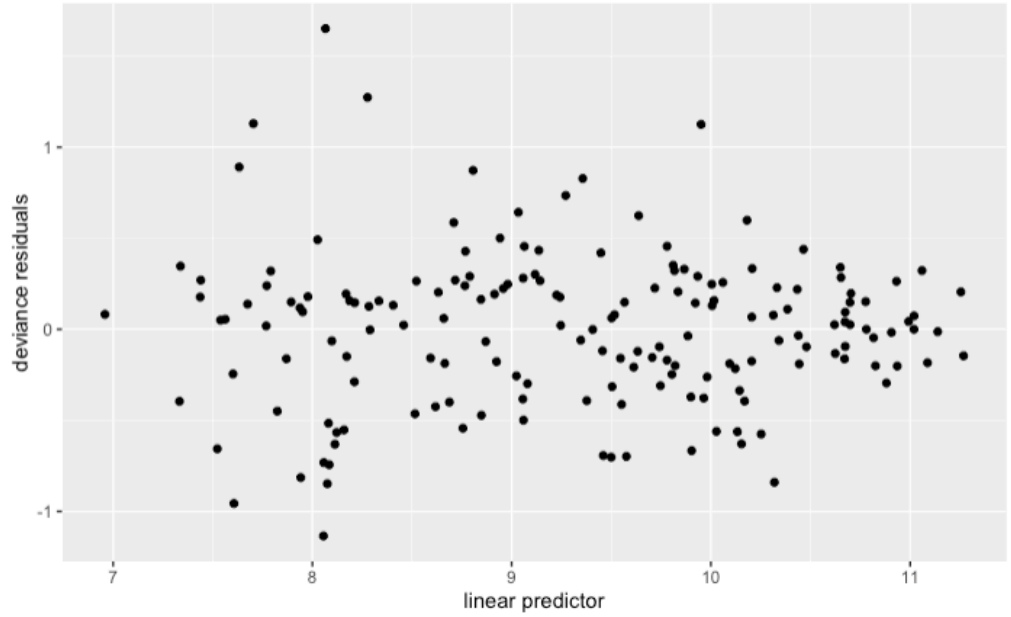
We start from a large model including all the predictor variables with interaction terms and do a sequential backward search using the step function to find the best-fitted model. The best-fitted model selected with interaction terms is:

```
log(GNIpercapita) ~ birthrate + deathrate + infantdeaths + life.exp.avg +
    life.exp.diff + region + birthrate:deathrate + birthrate:life.exp.avg +
    birthrate:life.exp.diff + deathrate:infantdeaths + deathrate:life.exp.avg +
    deathrate:life.exp.diff + infantdeaths:life.exp.avg + infantdeaths:life.exp.diff +
    infantdeaths:region + life.exp.avg:region
```

And an adjusted R-squared of 0.8374 indicates an improvement in predictability.

4.2.2 Outliers and Influential Points

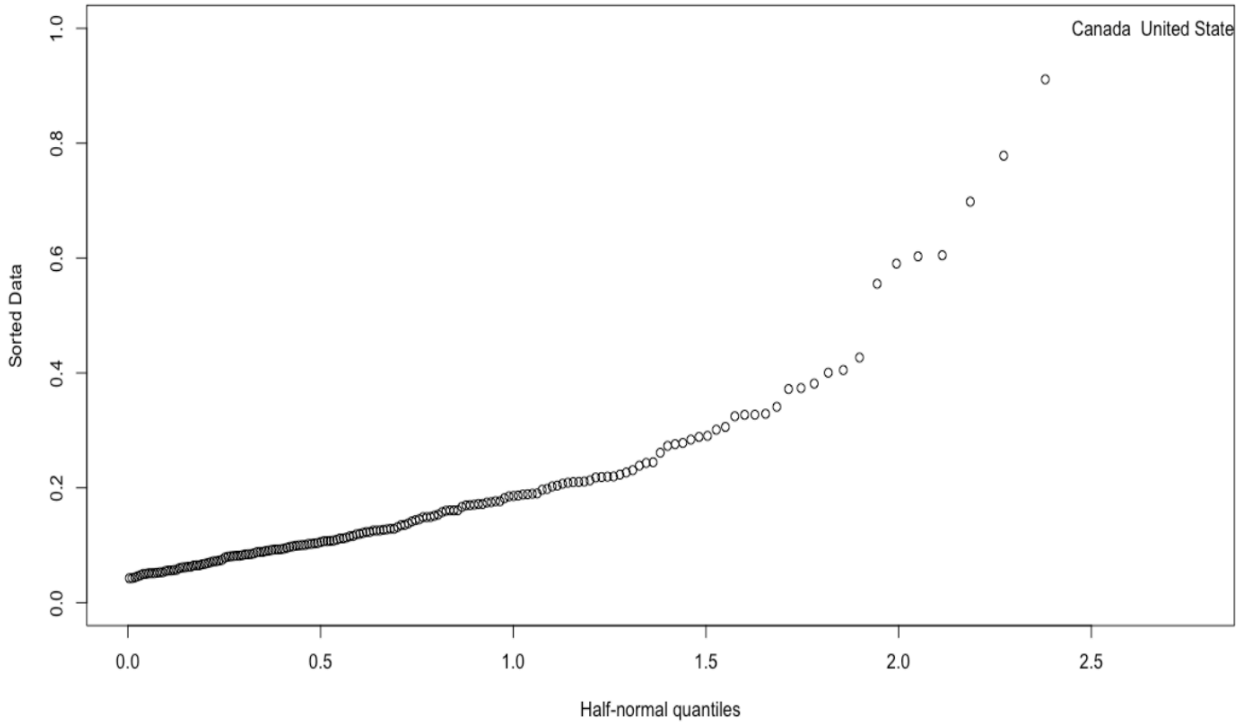
The plot of residuals vs. fitted values shows no noticeable pattern. The point above y=1.5 looks like a potential outlier.



The country is identified as Equatorial Guinea. However, there seems to be no problem with this data. It is just a little far away from others.

birthrate	deathrate	lifeexpM	lifeexpF	infantdeaths	GNIpercapita	region	country
33.73	9.54	57.06	59.26	64.5	16604.36	6	Equatorial Guinea

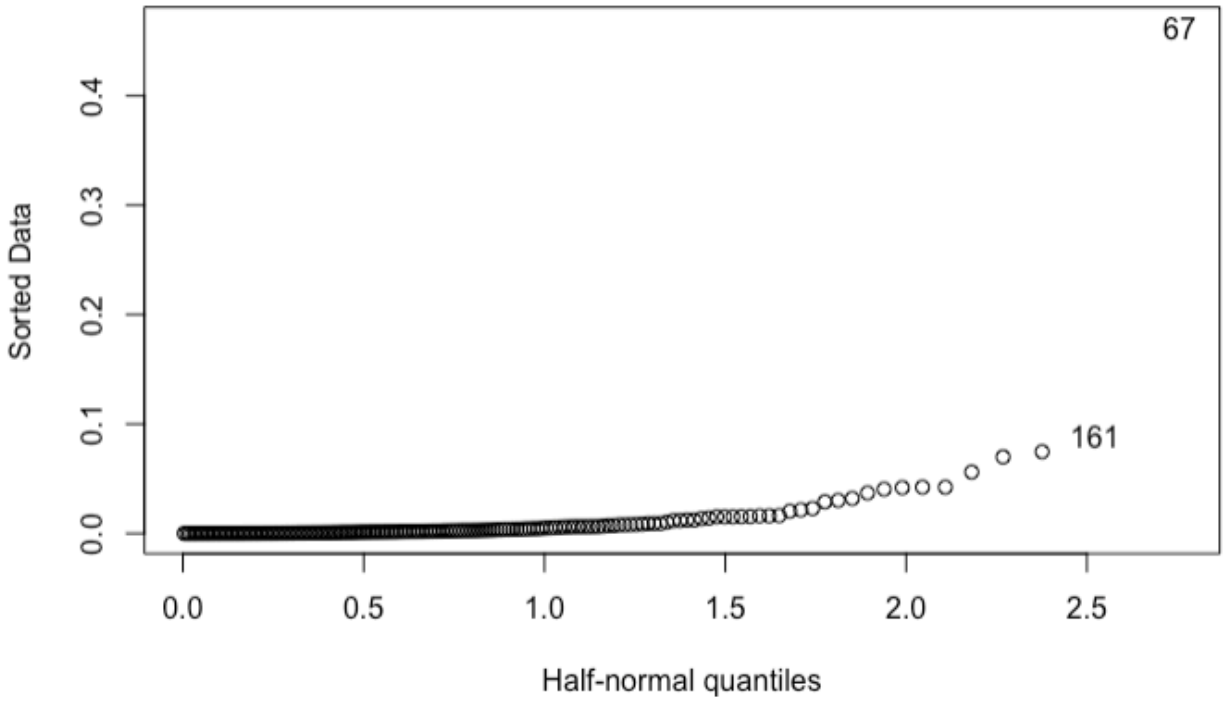
The half-normal plot shows that there are two potential outliers—Canada and the United States.



Canada and the United States are potential outliers because they are the only two countries in the North America region and both have high GNI per capita. But other regions have countries with both high and low GNI per capita. After removing two potential outliers, the adjusted R-squared decreases from 0.8374 to 0.8355. Thus, removing the two potential outliers does not improve our model.

We also plot the Cook distance against half-normal quantiles to reveal high influential points. Two high influential points are Haiti (161) and Turkmenistan (67). Turkmenistan is a high influential point because it has a much higher mortality rate of infants compared with countries that have similar GNI per capita.

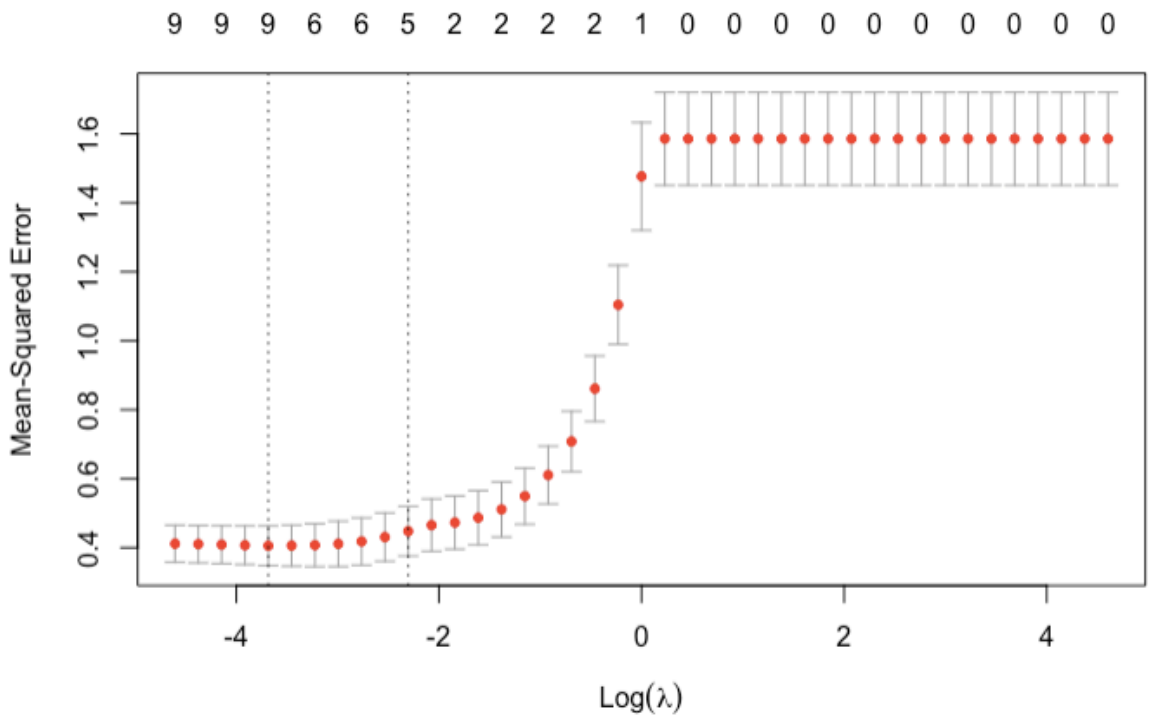
birthrate	deathrate	lifeexpM	lifeexpF	infantdeaths	GNIpercapita	region	country
19.38	5.1	69.67	75.62	10.6	13686.02	6	Libya
24.62	7.06	64.5	71.45	40.6	13615.49	3	Turkmenistan
14.6	5.8	70.13	75.22	20.4	13513.4	3	Azerbaijan



After removing two influential points, the adjusted R-squared improves from 0.8374 to 0.8436.

4.3 Model Selection by Lasso

We use the Lasso regression with cross-validation. We plot the cross-validation error curve. The plot shows that according to the log of lambda, the left dashed vertical line indicates that the log of the optimal value of log lambda is approximately -3.8, which is the one that minimizes the average AUC (area under the curve). The exact lambda value is 0.02511886. The cross-validation plot and corresponding model can be viewed as follow:




```
12 x 1 sparse Matrix of class "dgCMatrix"
                                        s0
(Intercept)                      3.846414583
deathrate                        0.059085629
infantdeaths                    -0.023996523
life.exp.avg                     0.075179816
life.exp.diff                   -0.003550681
regionEast Asia & Pacific        0.003092639
regionSouth Asia                -0.045886482
regionEurope & Central Asia      .
regionNorth America              0.514228677
regionLatin America             -0.106284355
regionMiddle East & North Africa 0.529791886
regionSub-Saharan Africa         .
```

We see that this model differs from the best-fitted model chosen by AIC by replacing the predictor birthrate by deathrate, infantdeaths, and life.exp.diff.

5. Conclusions

5.1 Summary of Findings

The GNI per capita can be well predicted by the indices of economic development included in this dataset, especially after a logarithmic transformation has been applied. The models we selected justify the use of the GNI per capita as a rough measure of the standard of living.

Adding interaction terms and removing high influential points improve adjusted R-squared since multicollinearity is a concern in the dataset.

5.2 Possible Weaknesses and Future Directions

Other economic data may also affect the GNI per capita that this dataset does not include. Examples are education level, inflation and exchange rates. There is much debate in Economics about the factors affecting the GNI per capita, but the factors discussed in this report are the most common ones that affect the GNI per capita.

There are also many other quantities for assessment of economic development and standard of living such as the Human Development Index (HDI), Genuine progress indicator (GPI), and OECD Better Life Index. It will be worthwhile to explore how different factors predict those indicators as well.

References

Data Catalog: 2018 Economic Indicators. (2018). The World Bank. Retrieved from https://datacatalog.worldbank.org/

Amadeo, K. (2020). *What Gross National Income Says About a Country.* The Balance. Retrieved from https://www.thebalance.com/gross-national-income-4020738

Cheng, M. (2020). *What Is Gross National Income (GNI)?* Investopedia. Retrieved from https://www.investopedia.com/terms/g/gross-national-income-gni.asp

Turkmenistan Has Highest Child Mortality Rate in Central Asia. (2019). turkmen.news. Retrieved from https://en.turkmen.news/news/turkmenistan-has-highest-child-mortality-rate-in-central-asia/

Hazuchova, N., & Savkova, J. (2017). *A Comparison of Living Standards Indicators.* European Journal of Business Science and Technology. 3. 10.11118/ejobsat.v3i1 .99.