World Bank Paper

Nick McCulloch with support from Cody Meagher and Stefano Musetti

2023-07-01

Introduction

For our project we selected the World Bank's Development Indicator's Database. The Database contains information by country, year, and topic, covering everything from healthcare to criminal justice and every year from 1960 to the present. We decided to explore the relationship between a country's education system and its economy.

The primary purpose of the World Bank is to support the development of lower income countries. However, although humanitarian in its aim, it's not a charity. The Bank offers loans and competitive grants and ideally hopes for a good rate of return. High impact and reliable development activities allow the bank to recoup its investment, and bad investments endanger its mission. Therefore, it's critical to identify which investments have the greatest relationship with the Bank's development goals and are most likely to pay off.

Education is often seen as a vehicle to better employment and prosperity. As students, we personally have made the decision to invest in education with the hope that the long-term reward is worth it. The government has taken an interest as well in the form of subsidized student loans and education grants. The potentially significant benefits of education are matched by significant expense, so it is critical to determine whether it is a good use of government, Bank, and student resources.

Initial Exploration

Our initial data set consisted of 10,113 observations of 65 variables. These include, Country and Series, two character-columns, and 63, year columns from 1960:2022. We first pivoted this data first longer, then wider, leaving us with 40 variables, Country and Year, and 38, Series variables which covered a mixture of education topics such as total adult literacy, primary, secondary, and tertiary education as a percentage of the population and economic measures such as GDP and GDP per capita. The following tables give an overview of the data.

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.2.3
```

```
wbdt_wide <- fread("wbdt_wide.csv")

a <- matrix(summary(wbdt_wide[,c("country","year","gnipc_constant", "edat_us_t", "literacy_at")]), ncol
colnames(a) <- c("country","year","gnipc_constant", "edat_us_t", "literacy_at")
a
```

```
##      country              year             gnipc_constant
## [1,] "Length:13671      " "Min.   :1960  " "Min.   :   189.4  "
## [2,] "Class :character  " "1st Qu.:1975  " "1st Qu.:  1501.7  "
## [3,] "Mode  :character  " "Median :1991  " "Median :  4002.3  "
## [4,] NA                   "Mean   :1991  " "Mean   : 11385.5  "
## [5,] NA                   "3rd Qu.:2007  " "3rd Qu.: 15040.2  "
## [6,] NA                   "Max.   :2022  " "Max.   :138198.5  "
## [7,] NA                   NA               "NA's   :8213  "
##      edat_us_t         literacy_at
## [1,] "Min.   : 0.20  " "Min.   :  5.405  "
## [2,] "1st Qu.:29.27  " "1st Qu.: 71.691  "
## [3,] "Median :50.33  " "Median : 90.721  "
## [4,] "Mean   :50.08  " "Mean   : 81.274  "
## [5,] "3rd Qu.:74.01  " "3rd Qu.: 96.451  "
## [6,] "Max.   :97.40  " "Max.   :100.000  "
## [7,] "NA's   :12495  " "NA's   :12601  "
```

challenges

The summary statistics above highlight some of the challenges that were found in the data set.

High Number of NA's: The data was overwhelmingly sparse. The year and country variables had no missing values, but every other variable was missing 70% or more of its values.

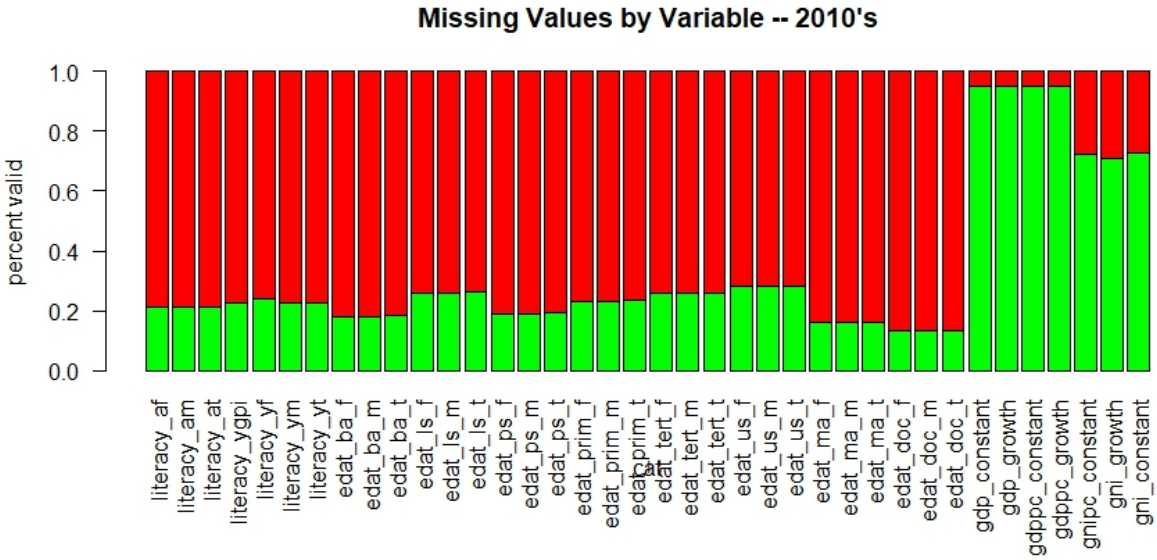


Figure 1: The 2010's had the lowest sparsity of any decade, but it was still extremely sparse.

Significant Outliers: A frequent finding in the project was the existence of massive global wealth inequality. At the high end some wealthy countries were a full 5 standard deviations from the mean. This was in part explained by subsequent analysis but the existence of both incredible riches and abject poverty was a notable aspect of the data and a complicating factor for statistical analysis, as the outliers had outsized weight in calculations and plots.

Plot 1 above highlights the significant outliers that both in the analysis and in that specific plot made interpretation difficult in the lower-income countries. Plot 2, significantly limits the max income shown

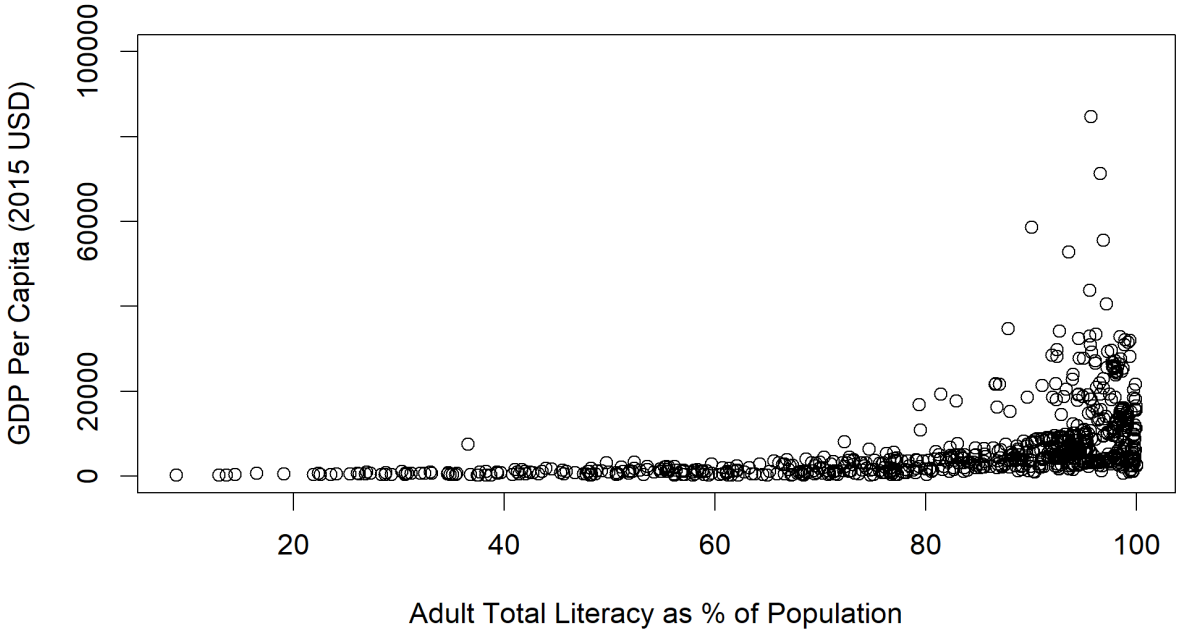


Figure 2: Plot1 highlights significant outliers

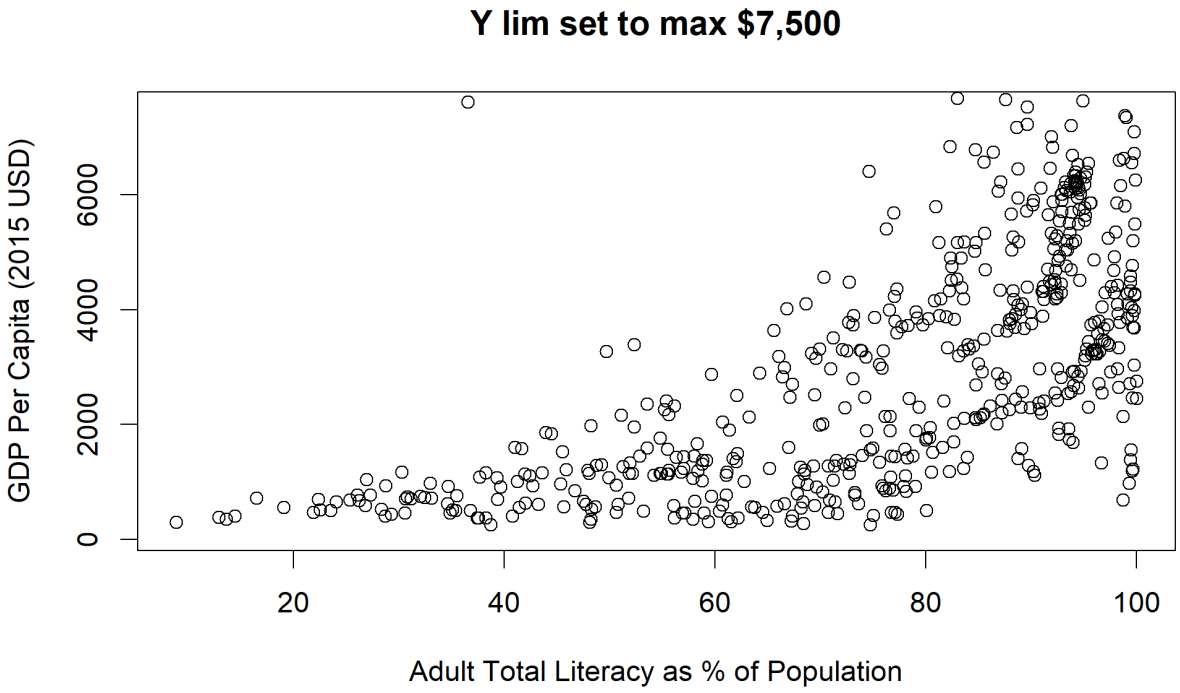


Figure 3: Plot 2 limits Y to $7,500 for legibility

which allows us to view a potential relationship between literacy and income. Many countries with high rates of literacy had low incomes but no country with literacy rates below 40% had a GDP per capita over $2,000. Based on these observations we were able to develop our initial hypothesis and plan for some of the issues to be accounted for in the analysis.

Formal Hypothesis

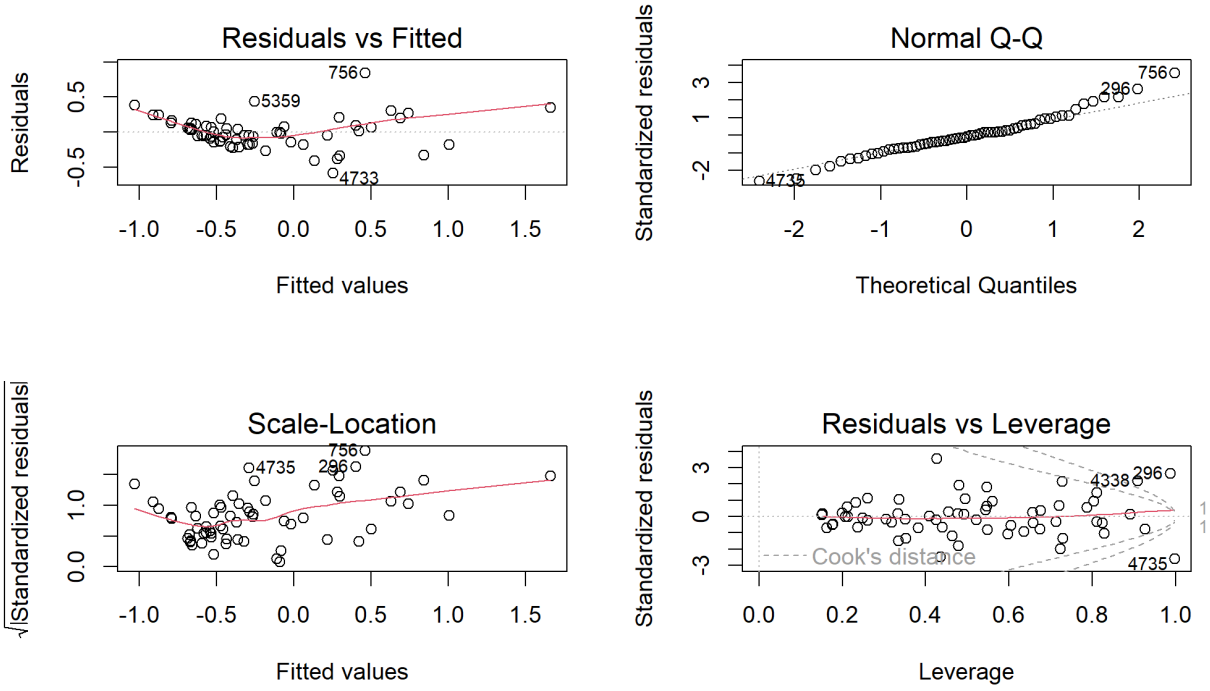
Based on our initial exploration we developed *our specific test hypothesis, that education did have a positive relationship with wealth.* This did relationship did not appear to be absolute. Certain education measures and certain brackets of country wealth seemed to be more or less correlated.

Analysis

We fit an initial model to examine which variables might be significant and look for other issues. Our initial model indicated that certain education measures might partially explain the variation in GDP per capita. However, previous plots caused us to suspect issues in the data.

Residuals

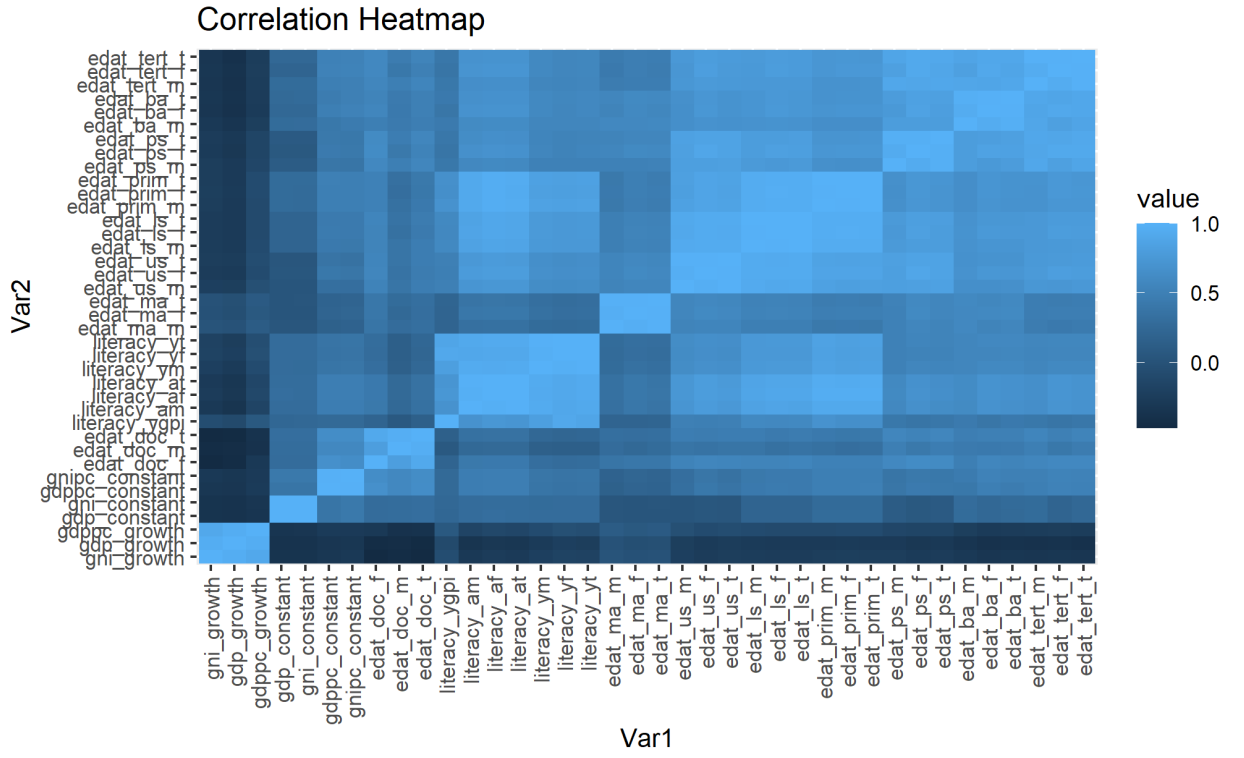
Upon examining our models, we found significant issues in the residuals.



The residuals/fitted plot showed large deviation from linearity to the right. The QQ plot showed light tails, indicating more data at the extremes compared to a normal QQ plot. The residuals vs leverage confirmed the presence of significant outliers. Finally the scale location plot was neither horizontal nor evenly spread, indicating heteroskedasticity, however, this wasn't conclusively confirmed by a BP test.

Multicollinearity

Many of the variables struck the team as inherently linked and unlikely to be independent. For instance, literacy and primary school attendance seemed unlikely to be independent. To confirm we created a correlation matrix and melted heat map. The heat map confirmed our suspicions.



Model Refinement

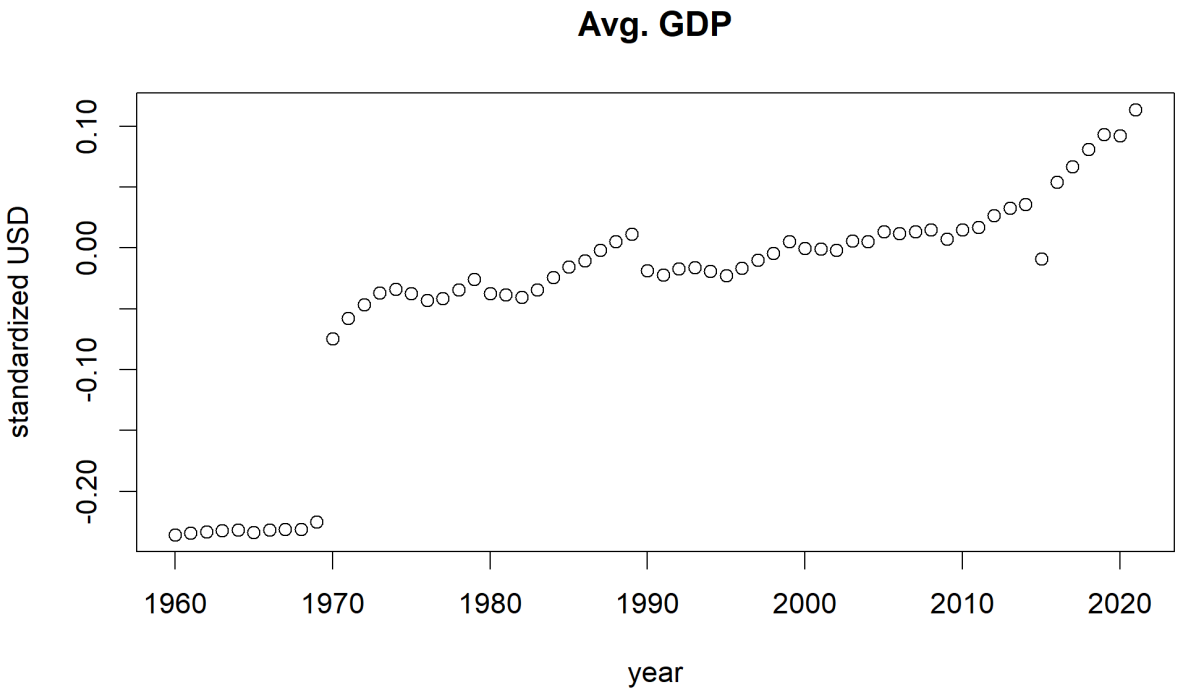
We refit our model based on our findings. Our new model, which used fewer predictors, excluded missing values, and accounted for some level outliers did improve on the original model. However, subsequent developments convinced us that significantly more work was needed.

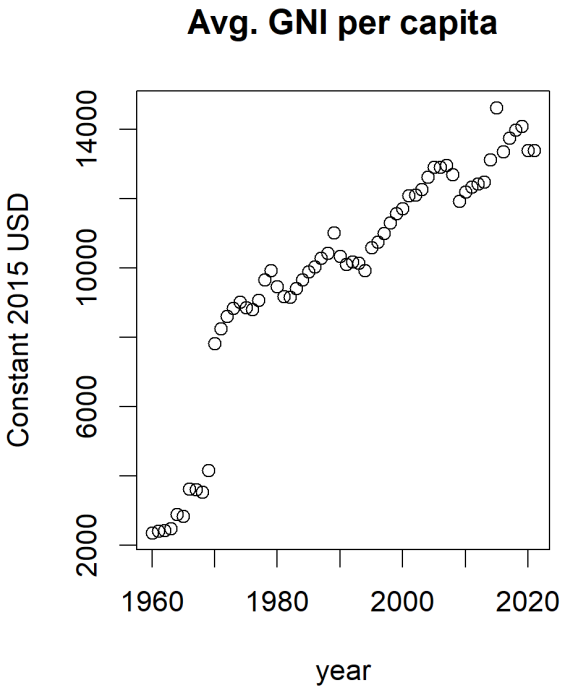
New Model Better but Still Inadequate

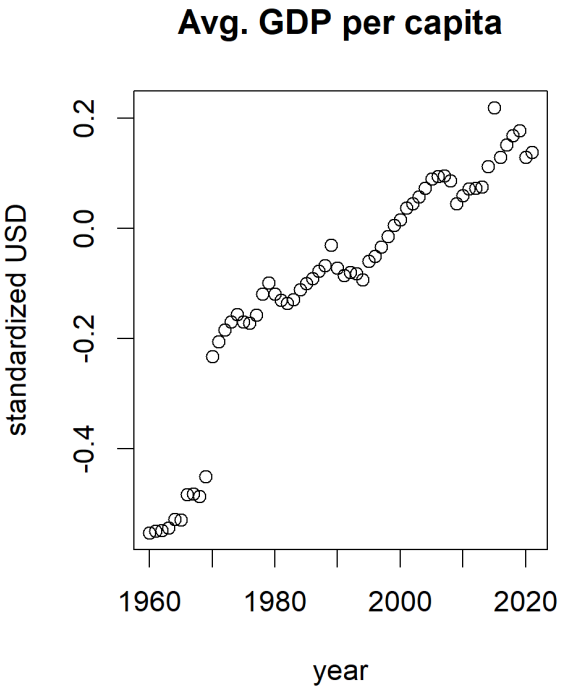
Our new model accounted for more of the variance than the previous model and we accounted for more potential issues. It's likely that further work in variable selection and clustering could pay dividends. However, any gains to be found there pale in comparison to incorporating time series analysis.

Time Series Analysis

We created new plots that explored the data from a time-series perspective. The results showed a strong and near-universal trend in economic measures over time.







The trend in economic improvement overtime was found in a majority of countries and was consistent across economic measures. It also explained variance more than any of the previous models. The only improvement we found was to explore economic measures based on the country itself.

Conclusion

Our conclusion was that education likely has some positive relationship with wealth. However, the relationship between pre-existing wealth and current wealth was much higher, and economic development proved to be much more dependent on time and fortune than could be accounted for in the relatively limited data available. If we were to do further research we would recommend enhanced variable selection, sparse data methods, and extensive time-series based analysis.