



JOINT SDG FUND

UN SDG Indicator Proxies

GR5055 Practicum Project 1

Columbia University, MA, Quantitative Methods in Social Sciences



Purpose

- Find or construct effective proxy measures to help address gaps in existing data
- Currently, there are large amounts of missing data, especially in recent years.
 - Filling data could assist in analysis and strategizing by the UN or Joint SDG Fund
- 6 target countries - Barbados, Cambodia, Indonesia, Malawi, Rwanda, Uzbekistan
 - Some analysis focuses on individual countries
- Several indicators of interest



Research and Literature Review

- The team began by conducting research into proxies that have been used elsewhere to inform which indicators would be beneficial to focus on
- Found potentially viable theoretical proxies for numerous indicators
 - Expenditure Share by Sector
 - Social Media Data
 - Climate Factors

A short horizontal bar with a teal segment on the left and an orange segment on the right.

Initial Data Exploration

- Given the ideas from the previous step, they performed initial searches for data relevant to the theory
- Many were found to be unfeasible
 - Indicator 5.b.1 - Proportion of Individuals who own a cellular phone, by sex
 - Little research available relevant to target countries
 - Data on theoretical proxies unavailable



Chosen Indicators

- Indicator 2.1.1
 - Prevalence of Undernourishment
- Indicator 2.2.2
 - Prevalence of Malnutrition among children under 5.
- Indicator 10.1.1
 - Growth Rate of the Income Share or Expenditure of the Bottom 40%



JOINT SDG FUND

Indicator 2.1.1

Prevalence of Undernourishment



Methodology

2.1.1



Methodology - SDG 2.1.1

- Focus on demand-side variables critical for undernourishment: food prices and poverty
- Calculate the “undernourishment poverty line” based on national-level staple food prices, food consumption patterns, energy requirements and expenditure structures
- Factor both the computed food expenditure poverty line and actual poverty rates into multivariate linear regression models for validation
 - Food expenditure relative to overall consumer-good price variations
 - Adjust for the translation from poverty rate to undernourishment rate



The Calculation of “Undernourishment Poverty Line”

Step 1-3: Adopted in regressions

1. Average food price

$$\text{price per calorie}_m = \frac{\sum_{i=1}^k (w_i \times P_{im})}{\sum_{i=1}^k (w_i \times C_i)}$$

2. Average calories required

$$MDER^E = \frac{\sum_{j=1}^n MDER(Age_j, Sex_j)}{n}$$

3. Food expenditure requirements

minimum required per capita food expenditure

$$= \sum_{m=1}^{12} (\text{Price per calorie}_m \times MDER^E)$$



The Calculation of “Undernourishment Poverty Line”

Step 4-5: Original plan unadopted in regressions (could be used for further study)

4. Total expenditure requirements (i.e. undernourishment poverty line)

$$\begin{aligned} & \text{minimum required per capita income} \\ &= \frac{\text{minimum required per capita food expenditure}}{\text{Engel Coefficient}} \end{aligned}$$

5. Estimation of the undernourishment rate based on the cumulative distribution of income

$$\begin{aligned} & \text{Prevalence of Undernourishment}^E \\ &= F(\text{minimum required per capita income}) \end{aligned}$$



Rationale for the Proxy Chosen - Research Support

- Martin Ravallion,
Does undernutrition respond to Income and Prices, Dominance Tests in Indonesia, 1992
 - The study in Indonesia over mid-1980s indicated staple food prices, along with average income levels, intra regional inequalities, have unambiguous effects on undernourishment
- Gustavo Anríquez, Silvio Daidone, Erdgin Mane,
Rising food prices and undernourishment: A cross-country inquiry, 2013
- Janice Meerman, Juliet Aphané,
Impact of high food price on nutrition, 2012



Rationale for the Proxy Chosen - Country Selection

Why Indonesia?

- Satisfy the first assumption behind this study's approach: people's accessibility to staple foods could be properly reflected by the national average market prices of the foods, calling for a nationwide food market that functions relatively well
- **Uzbekistan and Barbados:** undernourishment is less prevalent
- **Malawi and Rwanda:** aforementioned assumption is more likely to be violated due to higher level of market imperfections
- **Cambodia:** will be included in further study



Data

2.1.1



Indicator 2.1.1 Data

- Time Horizon: 2007-2019
 - Due to data availability of staple food prices and SDG 2.1.1 official data
- SDG 2.1.1 Dataset from UN SDG Indicators Database
 - Computed values due to difficulty in collecting household- or individual-level food consumption data
 - Computed estimates for all years based on the mean dietary energy consumption (DEC) and its coefficient of variation (CV) among population
 - Even less accurate during gap years between surveys
- Poverty rate data from BPS-Statistics Indonesia website
 - Rural, urban, and national rates
- Additional independent variable: percentage of population with sources of improved drinking water, from BPS-Statistics Indonesia website



- Datasets for calculation of “undernourishment poverty line”/inflation-adjusted food expenditure

| Dataset | Description | Source |
|--|---|--|
| Staple food prices | Monthly statistics of national average actual prices of several categories of staple foods | WFP data |
| Weight structure of food items | Weight structure of food items consumed by group with monthly expenditure of 200,000 to 299,999 rupiahs | BPS-Statistics Indonesia, 2019 |
| Food calorie content | Calorie content of food items included in the weighted basket | USDA National Nutrient Database for Standard Reference |
| Minimum Dietary Energy Requirements (MDER) | Minimum Dietary Energy Requirements (daily) by age, sex, height, weight and activity level | Human energy requirements: Report of a Joint FAO/WHO/UNU Expert Consultation |
| Age and sex structure + average height | Age and sex structure of population and average height of adult by gender (for calculation of MDER) | BPS-Statistics Indonesia, 2019; WorldData.info |
| Consumer Price Indices (CPI) | CPI (2007=100) for inflation adjustments of computed food expenditure minimum line | BPS-Statistics Indonesia |

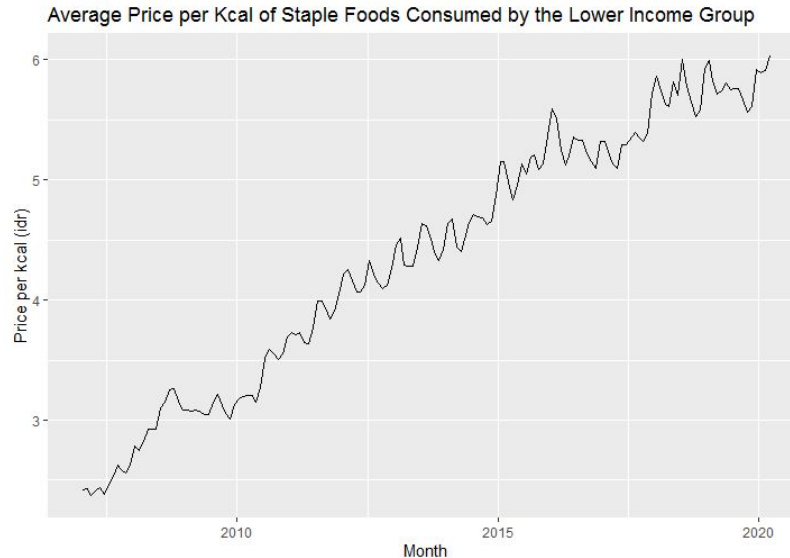


Results

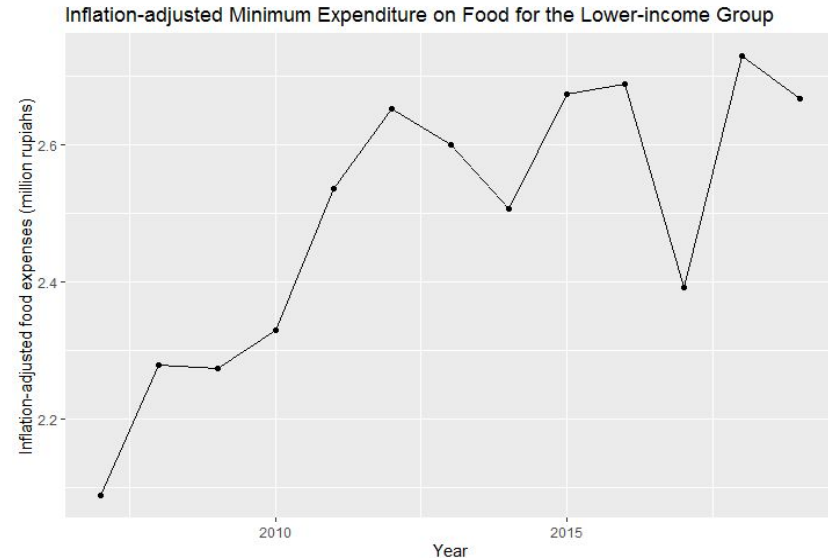
2.1.1

2.1.1 Results

- Apparent fluctuations in monthly data of staple food prices



- Overall increasing trend in inflation-adjusted minimum food expenditure faced by the lower-income group



2.1.1 Results

- Insignificant patterns found for inflation-adjusted food expenditure line and its interaction with rural poverty rates
- High R-square largely explained by poverty rates, especially rural poverty rate

| Dependent variable: | | | | |
|--------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | undernourishment | | | |
| | (1) | (2) | (3) | (4) |
| food_expenses | -1.933 (2.919) | -0.804 (2.893) | -0.782 (2.712) | |
| access_water | -0.001 (0.051) | -0.027 (0.052) | -0.028 (0.042) | -0.031 (0.042) |
| percent_poor | 1.809*** (0.373) | | | |
| rural_poor | | 1.596 (1.007) | 1.661*** (0.297) | 1.762*** (0.311) |
| urban_poor | | 0.095 (1.392) | | |
| rural_poor:food_expenses | | | | -0.045 (0.175) |
| Constant | -6.789 (11.645) | -11.409 (11.677) | -11.503 (10.937) | -13.142 (7.216) |
| Observations | 13 | 13 | 13 | 13 |
| R2 | 0.960 | 0.968 | 0.968 | 0.967 |
| Adjusted R2 | 0.946 | 0.951 | 0.957 | 0.957 |
| Residual Std. Error | 1.017 (df = 9) | 0.970 (df = 8) | 0.914 (df = 9) | 0.915 (df = 9) |
| F Statistic | 71.638*** (df = 3; 9) | 59.627*** (df = 4; 8) | 89.387*** (df = 3; 9) | 89.225*** (df = 3; 9) |
| Note: | | | | |

*p<0.1; **p<0.05; ***p<0.01



Limitations and Possible Further Study

- Limitations
 - Sample size too small (13 observations for 13 years)
 - Model depends on multiple strict assumptions
 - Price per Kcal should capture most supply-side variations in cultivation, stockbreeding, etc.
 - Equal access to food conditional on the same price
 - Basket of selected staple foods
- Further study with more available data
 - Region disaggregation (by dependent variable, food price, etc.)
 - Realizable for the original plan: survey data collected on the regional level (unavailable online)
 - Large potential for accuracy improvement
 - Rice subsidies: the “Raskin” program
 - Test on other countries, such as Cambodia



JOINT SDG FUND

Indicator 2.2.2

Prevalence of Malnutrition among children under 5

A horizontal bar with a teal segment on the left and an orange segment on the right.

Rationale

- This indicator was chosen due to its similarity to the SDG 2.1.1, where the difference is that the target group is narrowed down to children under 5.
- We have attempted many variables on SDG 2.1.1 but all hindered by the same problems of collecting a promptly dataset.
- Educational and climate data is a very great source to make approximations due to their availability and timeliness, however, they are not very related to SDG 2.1.1 but instead very relevant to SDG 2.2.2 when the scope narrows down to children under 5.
- Also, the great number of missing data as well as lags in reporting of SDG 2.2.2 make our work valuable to help the UN SDG Fund by providing alternative measures for this goal. We wish that this proxy can be used for future network analysis with other indicators of interest.



Methodology

2.2.2



Indicator 2.2.2 Methodology

- Multi-linear regression is used to validate our chosen proxies and see how closely they approximate the original measure of SDG 2.2.2. Based on the ANOVA test (p-value .001991), Model 3 was selected as the final model as it was statistically significant at the level of .05. The four models are:

Model 1: $\text{Goal2} \sim \text{Female_Literacy_Rate} + \text{Year}$

Model 2: $\text{Goal2} \sim \text{Female_Literacy_Rate} + \text{AVG_offset} + \text{Year}$

Model 3: $\text{Goal2} \sim \text{Female_Literacy_Rate} + \text{January} + \text{February} + \text{March} + \text{April} + \text{May} + \text{June} + \text{July} + \text{August} + \text{September} + \text{October} + \text{November} + \text{December} + \text{Year}$

Model 4: $\text{Goal2} \sim \text{Female_Literacy_Rate} + \text{offset_5y} + \text{offset_6y} + \text{offset_7y} + \text{offset_8y} + \text{offset_9y} + \text{offset_10y} + \text{January} + \text{February} + \text{March} + \text{April} + \text{May} + \text{June} + \text{July} + \text{August} + \text{September} + \text{October} + \text{November} + \text{December} + \text{Year}$



Data

2.2.2



Indicator 2.2.2 Data

- SDG 2.2.2 Dataset from the UN website
- Female Literacy Rate from the World Data Bank
- Primary School Female Enrollment Rate from the World Data Bank
- Climate Data (temperature deviation from the average of temperature from 1980-2005) by month, from <http://berkeleyearth.lbl.gov/country-list/>
- Countries of Interest: Low-income countries identified by the UN
- Time Horizon: 2000- 2019



Data Descriptions

Datasets for all countries were extracted and reorganized into a single csv file. The column names include:

- Country Name - The country's name
- Year - Year the data is from, from 2000 to 2019
- Month - Month of the year for the climate data
- Female_Literacy_Rate - Percent of the female population over the age of 15 who can read
- Female_School_Enrollment_5y_offset - Percent of girls enrolled in primary school 5 years prior to the year in question
- Female_School_Enrollment_6y_offset - see above, but based on data 6 years prior
- Female_School_Enrollment_7y_offset - see above
- Female_School_Enrollment_8y_offset - see above
- Female_School_Enrollment_9y_offset - see above
- Female_School_Enrollment_10y_offset - see above
- Female_School_Enrollment_AVG_offset - The arithmetic mean of the previous 6 columns
- Income_Group - The UN income classification of a country, including: Low Income, Middle low income, Middle high income, High income: non-OECD, and High income: OECD
- Region - The global region of the country
- UN_Member - 1 for countries that are members of the UN, 0 otherwise
- Climate_Data - The monthly standard deviation distance from the 1980-2005 mean temperature for a month\



Preprocessing of Data

- Linear imputation was used between data points to estimate data that otherwise would not have been usable. While this adds some complication, it should be valid as the data is in a time series format.
- The column “Climate_Data” has been categorized into monthly climate data. The 12 new columns represent each month of the year, and has been added to the target and educational data. Figure 2.2 shows the head of the final dataset that will be used for modeling and the validation process.



Results

2.2.2

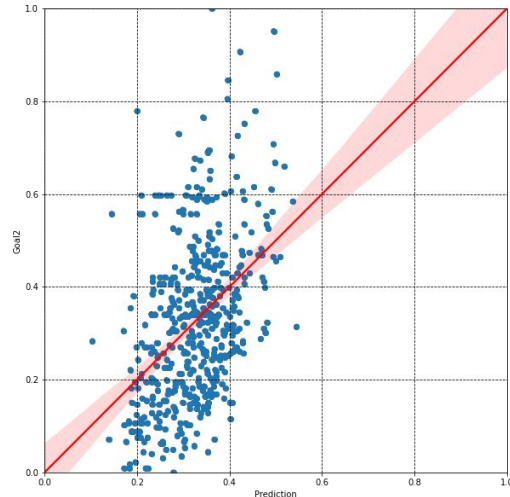


2.2.2 Results

Multi-linear regression is used to validate our chosen proxies and see how closely they approximate the original measure of SDG 2.2.2 for low-income countries.

R-squared = 0.187

Adjusted R-squared=0.164



OLS Regression Results

| | | | | | | |
|----------------------|------------------|---------------------|----------|-------|--------|--------|
| Dep. Variable: | Goal2 | R-squared: | 0.187 | | | |
| Model: | OLS | Adj. R-squared: | 0.164 | | | |
| Method: | Least Squares | F-statistic: | 8.017 | | | |
| Date: | Mon, 29 Nov 2021 | Prob (F-statistic): | 1.76e-15 | | | |
| Time: | 21:39:59 | Log-Likelihood: | 216.26 | | | |
| No. Observations: | 503 | AIC: | -402.5 | | | |
| Df Residuals: | 488 | BIC: | -339.2 | | | |
| Df Model: | 14 | | | | | |
| Covariance Type: | nonrobust | | | | | |
| ===== | | | | | | |
| | coef | std err | t | P> t | [0.025 | 0.975] |
| ----- | | | | | | |
| Intercept | 0.5997 | 0.074 | 8.092 | 0.000 | 0.454 | 0.745 |
| Female_Literacy_Rate | -0.1773 | 0.032 | -5.579 | 0.000 | -0.240 | -0.115 |
| January | -0.2540 | 0.085 | -2.996 | 0.003 | -0.421 | -0.087 |
| Febraury | 0.1417 | 0.073 | 1.950 | 0.052 | -0.001 | 0.284 |
| March | -0.0708 | 0.070 | -1.013 | 0.312 | -0.208 | 0.067 |
| April | 0.2503 | 0.063 | 3.980 | 0.000 | 0.127 | 0.374 |
| May | -0.0936 | 0.083 | -1.126 | 0.261 | -0.257 | 0.070 |
| June | -0.1615 | 0.080 | -2.007 | 0.045 | -0.320 | -0.003 |
| July | -0.1124 | 0.060 | -1.870 | 0.062 | -0.230 | 0.006 |
| August | 0.0352 | 0.071 | 0.494 | 0.621 | -0.105 | 0.175 |
| September | 0.1366 | 0.070 | 1.960 | 0.051 | -0.000 | 0.274 |
| October | -0.0039 | 0.082 | -0.048 | 0.962 | -0.164 | 0.156 |
| November | -0.0501 | 0.064 | -0.785 | 0.433 | -0.176 | 0.075 |
| December | -0.0719 | 0.055 | -1.312 | 0.190 | -0.180 | 0.036 |
| Year | -0.1011 | 0.030 | -3.335 | 0.001 | -0.161 | -0.042 |
| ===== | | | | | | |
| Omnibus: | 50.847 | Durbin-Watson: | 0.391 | | | |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 63.944 | | | |
| Skew: | 0.815 | Prob(JB): | 1.30e-14 | | | |
| Kurtosis: | 3.630 | Cond. No. | 29.2 | | | |
| ===== | | | | | | |

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

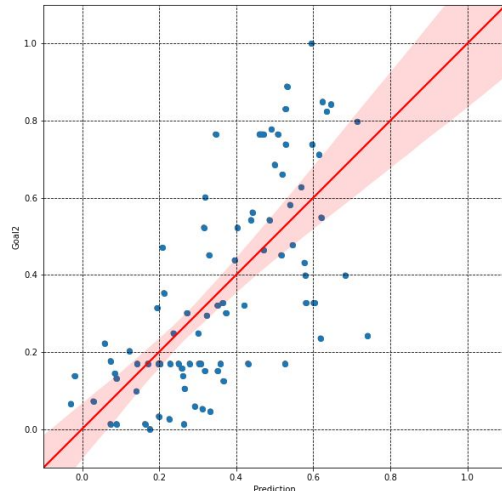


2.2.2 Results

Multi-linear regression is used to validate our chosen proxies and see how closely they approximate the original measure of SDG 2.2.2 for the 6 target countries of interest.

R-squared = 0.477

Adjusted R-squared=0.379



OLS Regression Results

```

=====
Dep. Variable:          Goal2      R-squared:                0.477
Model:                  OLS        Adj. R-squared:           0.379
Method:                 Least Squares      F-statistic:             4.885
Date:                   Mon, 29 Nov 2021    Prob (F-statistic):       2.58e-06
Time:                   22:01:16          Log-Likelihood:          20.301
No. Observations:       90              AIC:                    -10.60
Df Residuals:           75              BIC:                    26.90
Df Model:                14
Covariance Type:        nonrobust
=====

```

| | coef | std err | t | P> t | [0.025 | 0.975] |
|----------------------|---------|---------|--------|-------|--------|--------|
| Intercept | 0.9330 | 0.277 | 3.362 | 0.001 | 0.380 | 1.486 |
| Female_Literacy_Rate | 0.3254 | 0.080 | 4.063 | 0.000 | 0.166 | 0.485 |
| January | -0.2286 | 0.270 | -0.847 | 0.400 | -0.767 | 0.309 |
| February | -0.0190 | 0.201 | -0.094 | 0.925 | -0.420 | 0.382 |
| March | -0.3004 | 0.185 | -1.626 | 0.108 | -0.668 | 0.068 |
| April | 0.3122 | 0.262 | 1.193 | 0.237 | -0.209 | 0.834 |
| May | -0.1234 | 0.275 | -0.449 | 0.655 | -0.671 | 0.424 |
| June | -0.5131 | 0.264 | -1.941 | 0.056 | -1.040 | 0.013 |
| July | -0.1901 | 0.184 | -1.036 | 0.304 | -0.556 | 0.175 |
| August | -0.2081 | 0.138 | -1.504 | 0.137 | -0.484 | 0.068 |
| September | -0.1164 | 0.161 | -0.721 | 0.473 | -0.438 | 0.205 |
| October | -0.3495 | 0.201 | -1.735 | 0.087 | -0.751 | 0.052 |
| November | 0.0849 | 0.184 | 0.462 | 0.646 | -0.281 | 0.451 |
| December | 0.1851 | 0.207 | 0.895 | 0.373 | -0.227 | 0.597 |
| Year | -0.1153 | 0.098 | -1.171 | 0.245 | -0.312 | 0.081 |

```

=====
Omnibus:                1.122      Durbin-Watson:           1.015
Prob(Omnibus):           0.571      Jarque-Bera (JB):         1.002
Skew:                    0.002      Prob(JB):                 0.606
Kurtosis:                2.483      Cond. No.                  39.1
=====

```

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



Summary of Results

- The selected proxies are more readily available and have more complete datasets compared to the currently available dataset for SDG 2.2.2. In recent years, educational data for low-income countries generally and those 6 target countries in particular are becoming increasingly accessible, meaning this proxy has potential as an alternative indicator for SDG 2.2.2.
- Furthermore, the climate data is always up to date, meaning missingness is not an issue of concern, which was a major problem for the original measurement adopted by the UN.
- The wide availability and relative strength of the model as measured by the adjusted R-squared value provide justification to use this model.



Limitations and Possible Further Study

- Limitations
 - Incomplete datasets with many missing values, imputations may lead to biases.
 - Not all “low income countries” put great emphasis on agriculture, making the overall inferential analysis less significant than expected. However, results derived from the target countries are not bad.
- Possible further study
 - Further research to be conducted on specific months’ climate data, as we can see that some months have statistically significant patterns while the others are not. Since the 6 target countries rely heavily on agriculture, the crop production may be more likely to be affected by certain months.



JOINT SDG FUND

Indicator 10.1.1

Growth Rate of the Income Share or Expenditure of the Bottom 40%



Methodology

10.1.1



Methodology - SDG 10.1.1

- Data-based approach
- Test any possible variable that may be related to the income share of the bottom 40%
 - Initial tests performed on any data available to test for the existence of a relationship
 - Any countries for which data was available
 - Relationship tested based on correlation with Bottom 40% Income Share
 - Some of the theoretical variables discussed below
- Variables found to have a notable relationship with the indicator kept
 - Additional data sought
- Model built based on findings and optimized based on goodness-of-fit and predictions
- Linear Regression model built on 85% of data, tested on remaining 15%
 - Model built and tested on data for all countries first, then for 6 target countries
- If predictions were found to be successful, they can be used to calculate growth rate



Data

10.1.1

A horizontal bar with a teal segment on the left and an orange segment on the right.

Indicator 10.1.1 Data Sources

- UN Stats - official source of data regarding SDG Indicators
- World Bank, Development Research Group. Data are based on primary household survey data obtained from government statistical agencies and World Bank country departments. Data for high-income economies are from the Luxembourg Income Study database. <https://data.worldbank.org/indicator/>
- International trade center. Trade Map provides trade statistics and market access information for export development. It provides indicators on export performance, international demand, alternative markets and the role of competitors, and covers yearly trade data for 220 countries and territories and all 5,300 products of the Harmonized System. <https://www.intracen.org/>



Indicator 10.1.1 Data

- World Bank
 - Large amount of data for many variables
 - Several measures of correlation between each variable in the dataset and the Income Share of the Bottom 40% found
 - Those with consistent, strong relationships were selected to pursue more complete data
- Variables Chosen
 - Gini Coefficient (Our World in Data)
 - Education (Our World in Data)
 - Government Expenditures (Our World in Data)



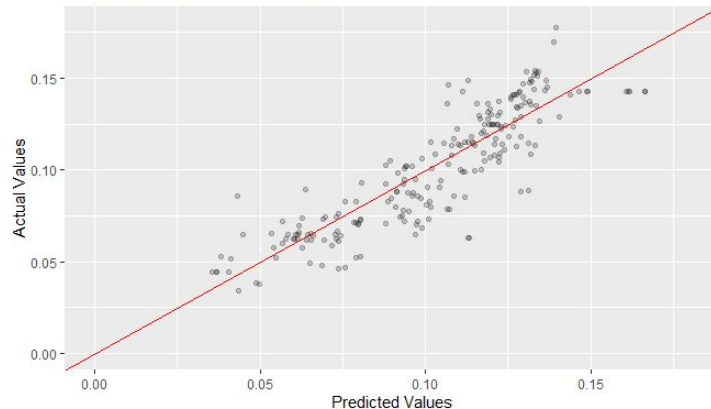
Results

10.1.1

10.1.1 Results

- Linear Regression model built on 85% of the data
 - $R^2 = .74$
 - Income Share of Bottom 40% as a function of Gini Coefficient, Average Number of years of Schooling, and Total Government Expenditures
 - Predictions compared to actual values of the income share of the bottom 40%

Predicted vs. Actual Values



Call:

```
lm(formula = Income_Share_Bottom_40 ~ Gini.index + SchoolYears + govexpend, data = train)
```

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|--|-----------|-----------|----------|----------|----------|
| | -0.057992 | -0.009505 | 0.000706 | 0.010049 | 0.057774 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|-------------|
| (Intercept) | 1.657e-01 | 3.870e-03 | 42.82 | < 2e-16 *** |
| Gini.index | -2.486e-03 | 6.443e-05 | -38.59 | < 2e-16 *** |
| SchoolYears | 1.190e-03 | 2.004e-04 | 5.94 | 3.7e-09 *** |
| govexpend | 6.267e-04 | 4.427e-05 | 14.16 | < 2e-16 *** |

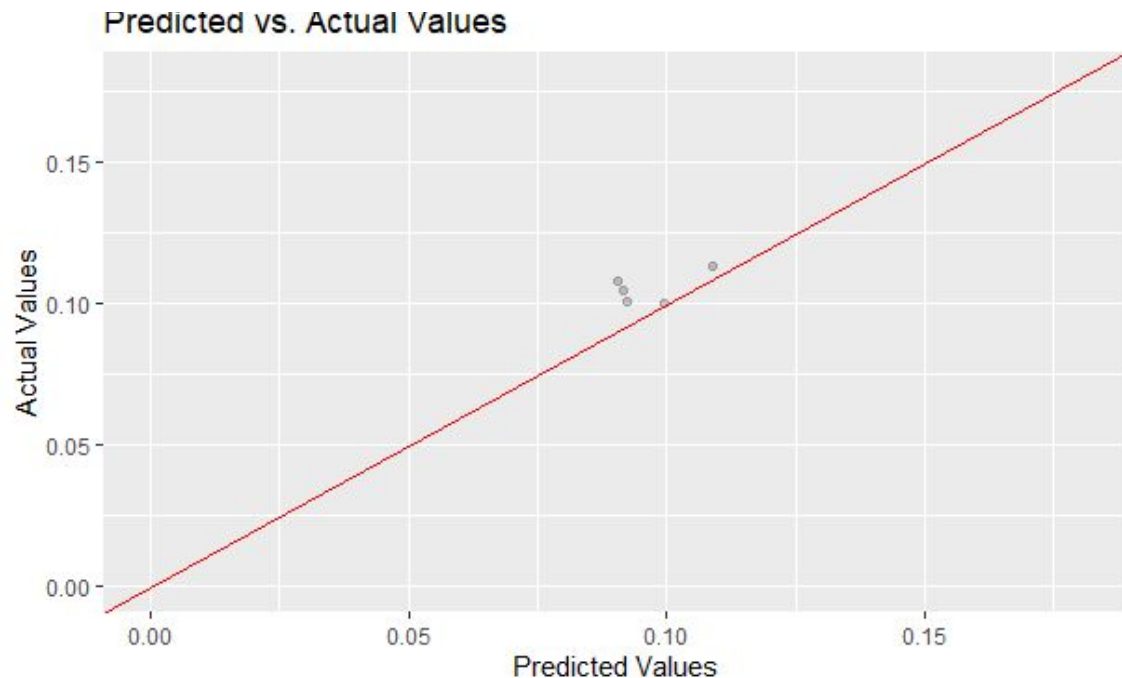
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01651 on 1236 degrees of freedom
Multiple R-squared: 0.7427, Adjusted R-squared: 0.7421
F-statistic: 1189 on 3 and 1236 DF, p-value: < 2.2e-16

Predictions for Target Countries

The number of predictions for the 6 target countries was small, due to missingness in the data. However, what remained were often fairly strong predictions.

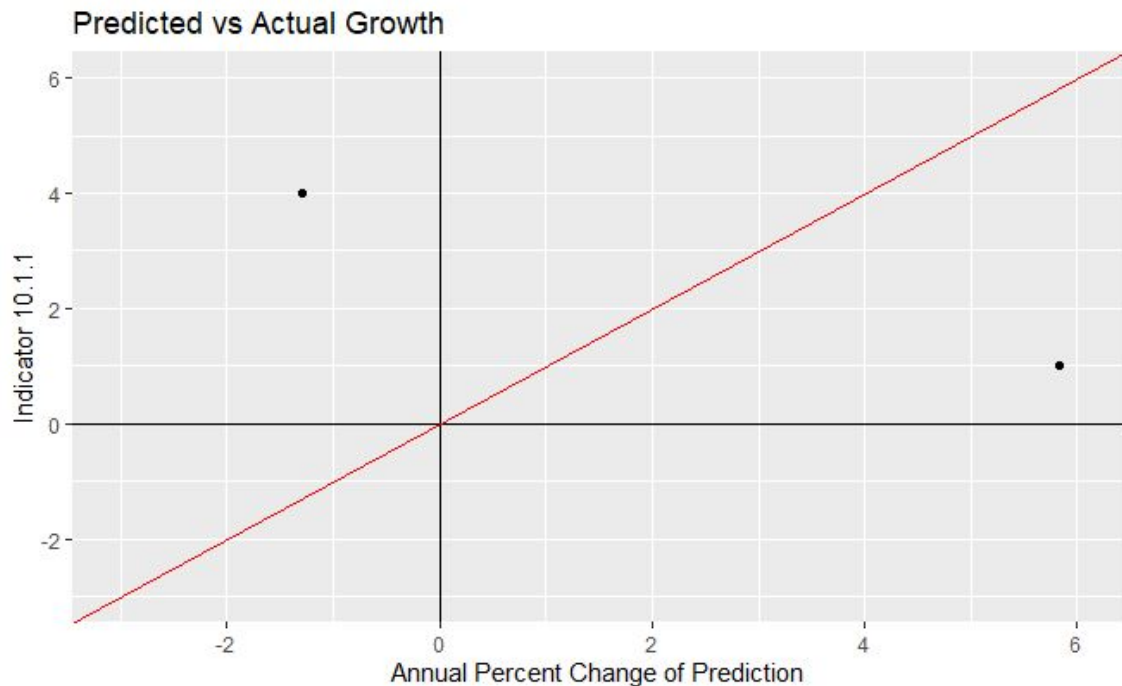
$$2 < n < 6$$



Predicted Growth and Official Data

Few observations existed for the same country in consecutive years, especially when combined with the missingness of the official indicator data.

$$0 < n < 5$$





10.1.1 Discussion - Strengths

- Model's predictions were fairly strong, subject to the precision required, using only three common variables.
- Model could be improved in accuracy with additional data if obtainable
- The predictions could theoretically be used to fill in some of the gaps of missing data, providing potential use to the Joint SDG Fund, with some caveats


A short horizontal bar with a teal segment on the left and an orange segment on the right.

Limitations and Possible Further Study

- Additional variables
 - Family size/planning
 - Child Employment
 - More specific measurement of social program coverage
- Low Sample Size
 - Few explanatory variables
 - Extrapolation
 - Model built on disproportionate amount of data from developed countries
 - Time-lag not successfully addressed in full



Indicator 10.1.1 Appendix - Some Tested Variables

|  Variable | Correlation with Income share of Bottom 40% |
|--|---|
| Proportion of Consumption on Food and Accommodation | -0.135095 |
| Proportion of Consumption on Clothing | -0.2754005 |
| Loan Default Rate | 0.547835 |
| Expenditure on Durable Goods | -0.9360472 |
| Number of Commercial Bank Branches | 0.2352157 |
| Number of McDonald's | -0.05808015 |
| Incarceration Rate | -0.182441 |
| Trade Data | Range: (.001 to .21 for all countries, .31 to .83 for target countries) |