Exploring How Income Inequality is Associated with Economic and Social Factors

Final report for Data Analysis and Interpretation Specialization
December 11, 2016

Introduction

The purpose of this study is to identify economic and social factors associated with income inequality, as measured by Income Share – the distribution of income to the richest 10 percent of earners on a country by country basis. A wide range of predictors has been evaluated, including Income Per Capita, GDP Per Person, Internet Use, Electricity Access, Fuel Access, C02 Emissions, Unemployment, Labor Participation, Migrant Population, Urban Population, Savings, Business Startup Costs, Business Startup Time, and Tax Rates.

A slow recovery following the 2008 Great Depression has highlighted a gap in income inequality that has been widening for several decades. While jobs and the economy have grown at a moderate pace in recent years, many middle- and lower-income workers have seen earnings flatten or drop. Identifying factors associated with income inequality will provide insights on how to foster income growth for the middle- and lower- class.

With insights on what drives income inequality, voters, media outlets, government officials and politicians will have a better framework to discuss, debate and plan policies. Such perspective is desperately needed to bypass heated and vacuous exchanges as seen in the recent U.S. presidential election. Instead, informed discussions could help lead to a healthier global economy, and meet a World Bank Group's goal to promote shared prosperity in every country.

Methods

Sample

The data comes from the World Bank's primary collection of development indicators, representing the most current and accurate information from official sources. The full dataset includes (N=264) countries and 1,446 variables from 1960 to 2016. For this study, a subset (N=102) was created to include only countries with values reported for all 14 predictor variables and the target – share of total income for top earners – in one or more years from 2005 to 2014. Income distribution was calculated from household surveys when available, or otherwise estimated from grouped data, and adjusted for household size to reflect income per capita.

Measures

The target – Income Share – and predictor variables are all continuous quantitative values. To ensure an adequate sample size, each was averaged over a 10-year period, 2005-2014, while accounting for years with missing data. For example, if a variable had data for only three out of those five years, then the values were summed and divided by just three.

Here is the target and the predictor variables:

- Income Share: Income share held by highest 10%
- Income Per Capita: Adjusted net national income per capita (current US$)
- GDP Per Person: GDP per person employed (constant 2011 PPP* $)
- Internet Use: Internet users (per 100 people)
- Electricity Access: Access to electricity (% of population)
- Fuel Access: Access to non-solid fuel (% of population)
- C02 Emissions: KG per PPP* $ of GDP
- Unemployment: % of total labor force (modeled ILO** estimate)
- Labor Participation: Labor force participation rate, total (% of total population ages 15-64) (modeled ILO** estimate)
- Migrant Population: International migrant stock (% of population)
- Urban Population: Urban population (% of total)
- Savings: Gross savings (% of GDP)
- Business Startup Costs: Cost of business start-up procedures (% of GNI*** per capita)
- Business Startup Time: Time required to start a business (days)
- Tax Rates: Taxes on income, profits and capital gains (% of total taxes)

*PPP is purchasing power parity, **ILO is International Labor Organization, ***GNI is Gross Domestic Income

Analyses

Mean, standard deviation and minimum and maximum values were calculated for the target – Income Share – and each predictor variable to examine their distributions. Pearson correlation and scatterplots were used to test bivariate associations between the target and predictors.

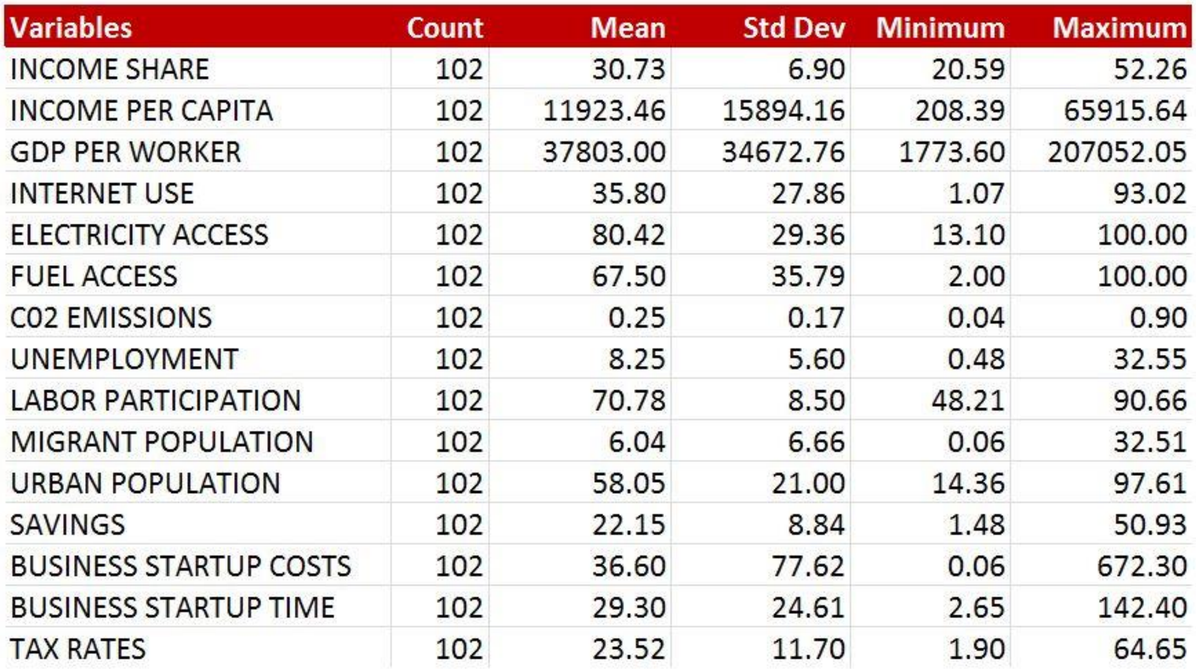
A lasso analysis using the least-angle regression algorithm was performed to identify a subset of predictor variables best associated with Income Share. The change in mean-squared error at each step of a cross-validation process, using 10 folds, was used to select the best predictors for the model. Prior to analysis, all predictor variables were standardized to ensure each had a mean of 0 and a standard deviation of 1. Because the sample is small (N=102), the data was not split into training and testing groups.

Results

Descriptive Statistics

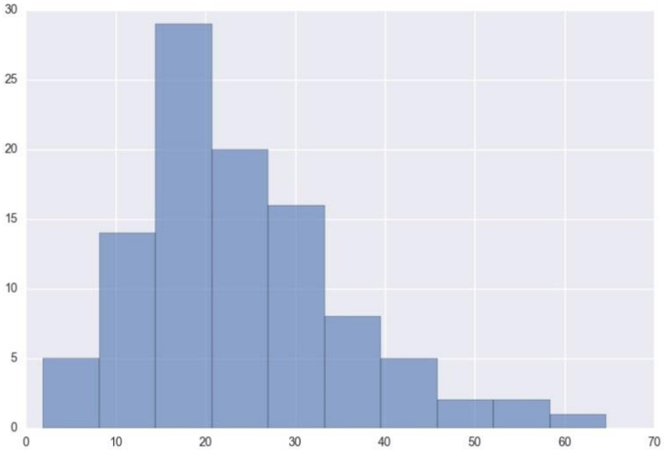
Table 1 shows mean and spread statistics for all variables. Average share of total income held by the highest 10% of earners in countries - the target variable - is 30.7% (sd=6.90), with a minimum of 20.6% and a maximum of 52.3%. Figure 1 shows that observations of Income Share have a unimodal distribution that is right skewed.

Table 1: Descriptive Statistics



Variables	Count	Mean	Std Dev	Minimum	Maximum
INCOME SHARE	102	30.73	6.90	20.59	52.26
INCOME PER CAPITA	102	11923.46	15894.16	208.39	65915.64
GDP PER WORKER	102	37803.00	34672.76	1773.60	207052.05
INTERNET USE	102	35.80	27.86	1.07	93.02
ELECTRICITY ACCESS	102	80.42	29.36	13.10	100.00
FUEL ACCESS	102	67.50	35.79	2.00	100.00
C02 EMISSIONS	102	0.25	0.17	0.04	0.90
UNEMPLOYMENT	102	8.25	5.60	0.48	32.55
LABOR PARTICIPATION	102	70.78	8.50	48.21	90.66
MIGRANT POPULATION	102	6.04	6.66	0.06	32.51
URBAN POPULATION	102	58.05	21.00	14.36	97.61
SAVINGS	102	22.15	8.84	1.48	50.93
BUSINESS STARTUP COSTS	102	36.60	77.62	0.06	672.30
BUSINESS STARTUP TIME	102	29.30	24.61	2.65	142.40
TAX RATES	102	23.52	11.70	1.90	64.65

Figure 1: Histogram for Income Share



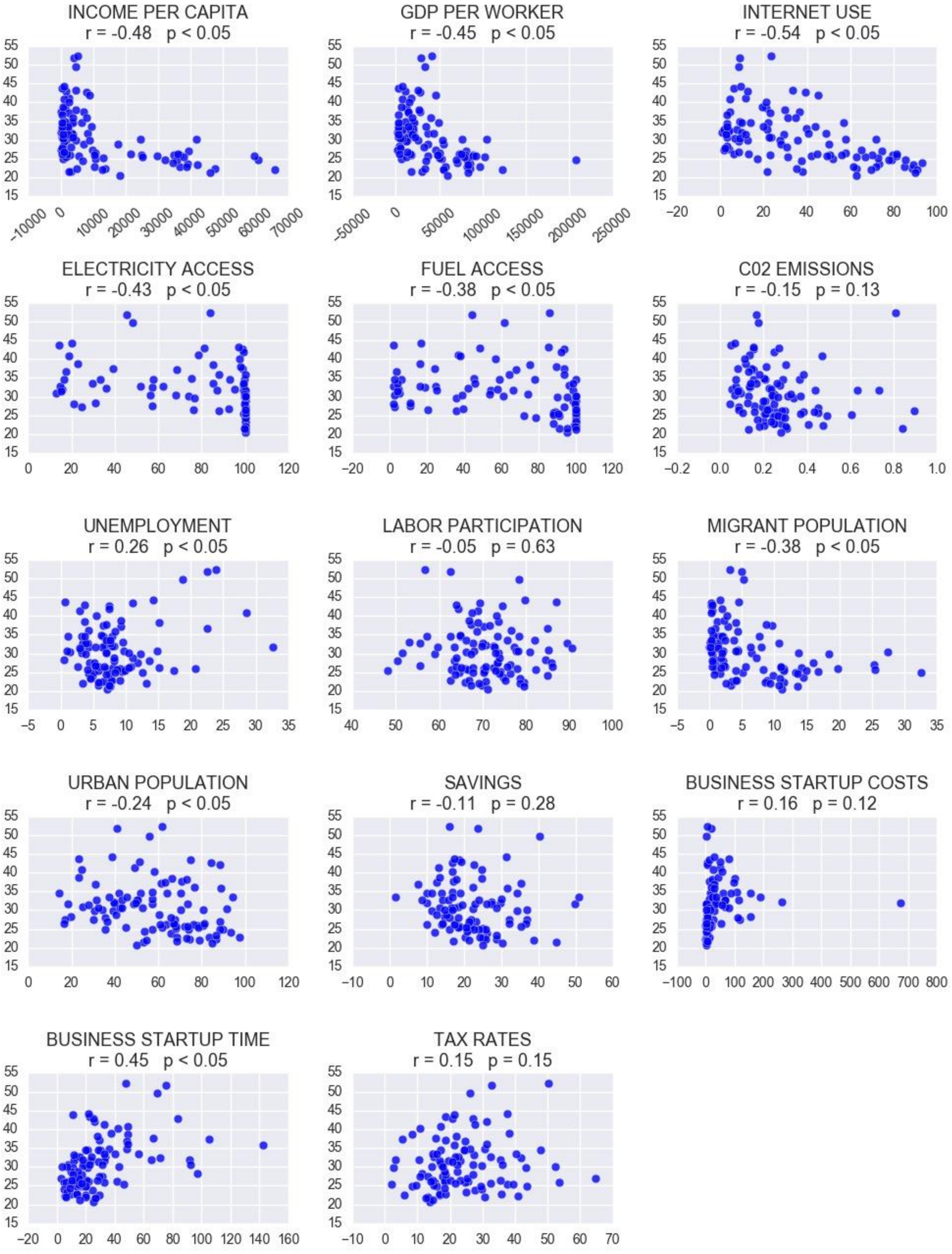
Bivariate Analyses

Scatter plots, along with Pearson correlation coefficients (r) and p-values (p), show linear associations between Income Share and each predictor (Figure 2 on next page). The data shows that Income Share, or we can say income inequality, is:

- Moderately higher when Business Startup Time is longer
- Somewhat higher when Unemployment is higher
- Moderately lower when Income Per Capita, GDP Per Worker, Internet Use and Electricity Access is higher
- Somewhat lower when Fuel Access, Urban Population and Migrant Population is higher
- Not significantly associated with C02 Emissions, Labor Participation, Savings, Business Startup Costs and Tax Rates

The scatter plots also show that many relationships are curved rather than straight, with data often skewed to the right. Some plots, such as GDP Per Worker and Business Startup Costs, and possibly C02 Emissions and Business Startup Time, indicate potential outliers.

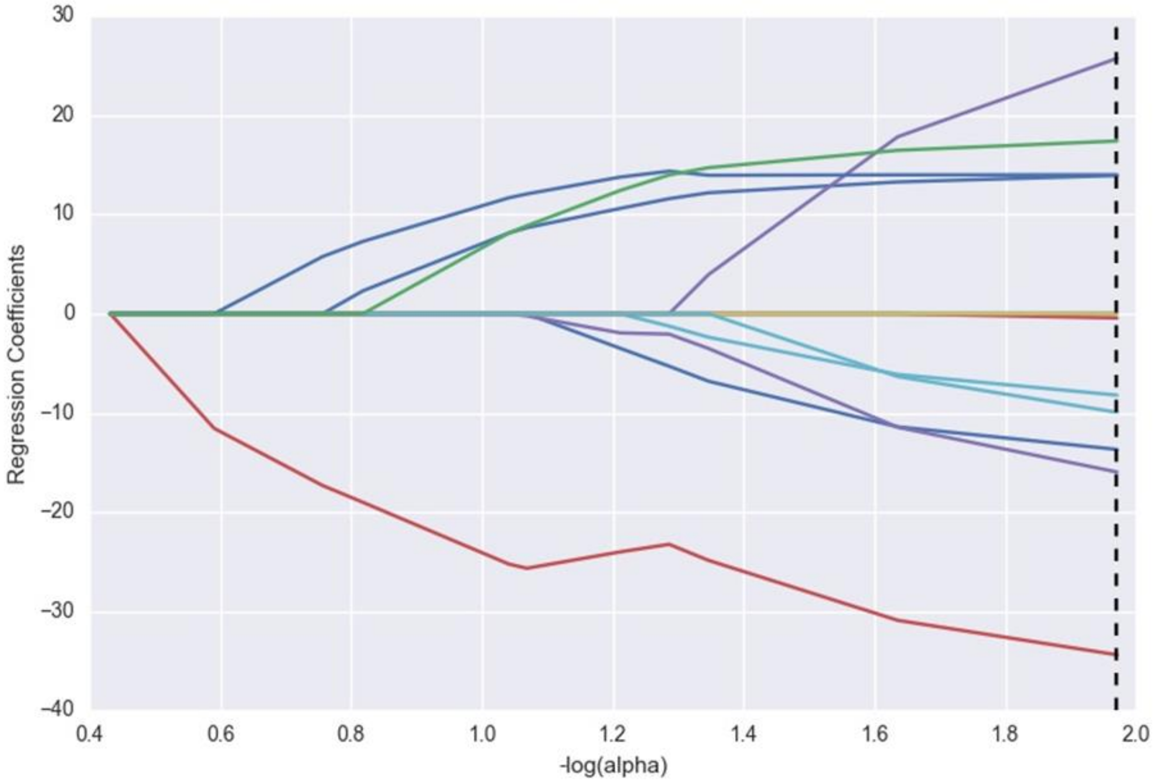
Figure 2: Linear associations Between Income Share and Predictors



Lasso Regression Analysis

Figure 3 shows the relative importance of predictor variables during the lasso regression selection process, including where they were added and how the regression coefficients changed at each step. The model retained 10 predictors, though Migrant Population was close to 0, while GDP Per Worker, Fuel Access, Labor Participation and Savings were excluded. The 10 selected predictors together account for 58.6% of the variance in Income Share.

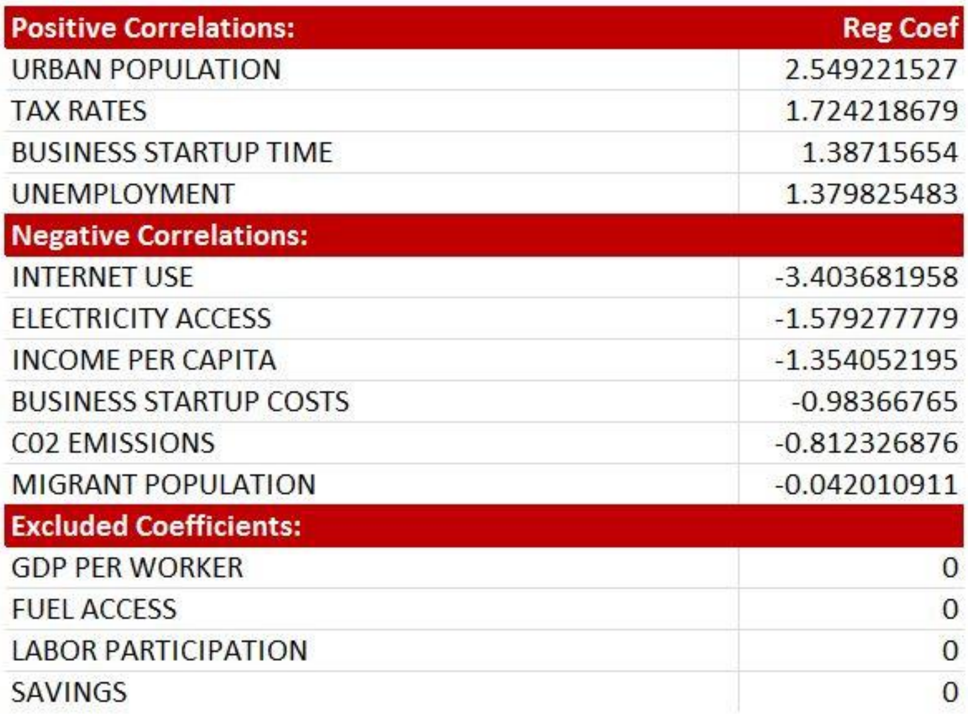
Figure 3: Regression Coefficients for Lasso Paths



The regression coefficients for the predictor variables (Table 2) show that:

- Higher levels of income inequality are most strongly associated with higher Urban Population, followed by higher Tax Rates, Business Startup Time and Unemployment
- Lower levels of income inequality are most strongly associated with higher Internet Use, followed by higher Electricity Access, Income Per Capita, Business Startup Costs, C02 Emissions and Migrant Population

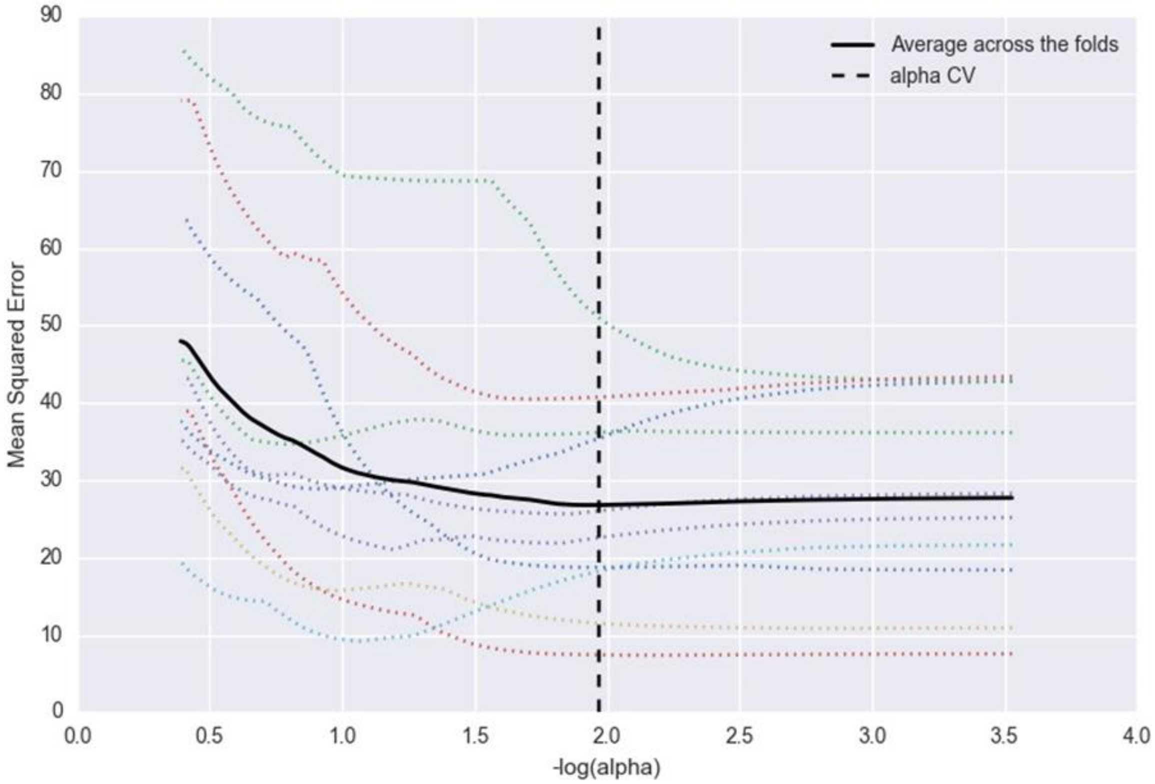
Table 2: Lasso Regression Coefficients for Predictors



Positive Correlations:	Reg Coef
URBAN POPULATION	2.549221527
TAX RATES	1.724218679
BUSINESS STARTUP TIME	1.38715654
UNEMPLOYMENT	1.379825483
Negative Correlations:	
INTERNET USE	-3.403681958
ELECTRICITY ACCESS	-1.579277779
INCOME PER CAPITA	-1.354052195
BUSINESS STARTUP COSTS	-0.98366765
C02 EMISSIONS	-0.812326876
MIGRANT POPULATION	-0.042010911
Excluded Coefficients:	
GDP PER WORKER	0
FUEL ACCESS	0
LABOR PARTICIPATION	0
SAVINGS	0

Figure 4 shows the mean square error for changes in the penalty parameter at each step in the selection of predictors. While there is variability across individual cross-validation folds, the mean follows a similar path, decreasing and then levelling off as expected.

Figure 4: Mean square error for each fold



Conclusions

Overview

This study uses a lasso regression analysis to select variables best associated with income inequality in countries (N=102) from 2005 to 2014, as tracked by the World Bank's primary collection of development indicators. Income inequality is measured as total share of income held by the richest 10% of earners in each country. The data shows income share averages 30.7%, and ranges from 20.6% to 52.3%, which is a significant spread.

Out of 14 predictor variables examined, the analysis retained 10, which together account for 58.6% of the variability in income share. Higher levels of income inequality are most strongly associated with larger urban populations, while lower levels are mostly linked to more use of the Internet.

From an economic perspective, income share is more lopsided when tax rates are higher and business startup times longer, but flatter when income per capita, business startup costs and C02 emissions are higher. Also, income is more unbalanced when unemployment is higher, but more even when electricity access is better and migrant populations larger.

Implications

Results show that business climates may play a role in how much wealth flows to or remains in the hands of the rich. When tax rates are high and entrepreneurs face barriers starting businesses, income inequality is higher. These factors, also associated with high unemployment, might limit opportunities and slow economic growth.

However, when earnings, startup costs and air pollution are higher, income is distributed more evenly. These factors could be byproducts of robust investments and business activities. In addition, key resources such as the Internet and electricity may be important tools to spark the economy and help level incomes.

To promote the World Bank's goal of shared prosperity in every country, this study provides context for discussion and debate, and points to areas where we can look deeper, such as impacts from large urban areas. Also, some results should give pause to recent political rhetoric. For example, while higher unemployment is associated with higher income inequality, a larger migrant population is not.

Limitations and Suggestions

This study's data sample is small (N=102), and focused on a 10-year average out of 57 years available in the World Bank's collection of development indicators.

Recommendation: Build a stronger model by seeking additional data as the World Bank updates its collection, and consider how to add more observations, such as other ways to measure income inequality and including more years.

While this study examined a target variable and 14 predictors, the full dataset contains a total of 1,446 features. Other factors, not looked at here, may be more closely associated with income inequality, and some could confound the predictors studied.

Recommendation: Widen the study's scope by conducting analysis of additional predictors, including binning or building composite values where it makes sense.