Savings Rate Analysis

team undecided

Yutong Liang, Jaja Sothanaphan, Zhehui Liang, Miku Suga

CSCI 1951a: Data Science

Problem Statement

Hypothesis Testing:

We are going to see if there is a significant correlation between a general macroeconomic phenomena, e.g. unemployment rate, and the savings rate, the ratio of savings to nominal GDP of the country.

With following macroeconomic variables as our independent variables:

Unemployment rate, in %: Ratio of unemployed population to labor force

Male-to-population ratio, in %: Ratio of male population to the overall population

Female-to-population ratio, in %: Ratio of female population to the overall population

Total population

Population density: Number of people per square kilometers of land area

Nominal GDP: in current local currency units relative to the U.S. dollar

Nominal GDP per capita, in current local currency units relative to the U.S dollar

GDP growth, in %: Change in GDP after an year

Age dependency ratio of young working age, in %: Ratio of working age to non working age population

Literacy rate, in %: Ratio of number of literate people to total population (for age > 15)

Education, in %: Ratio of number of people with educational attainment of Doctoral or equivalent

Motivation

Savings rate is an interesting component to investigate when observing the economy of a country, e.g. too low savings rate can be a sign of trouble.

We are **not** doing a prediction model because 1. Measuring the success of our prediction model is extremely hard since there is no "correct" data 2. A prediction model without initially knowing which features are correlated would not be a good start.

Data

Data Source

https://data.worldbank.org
https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-people.html
https://data.oecd.org/hha/household-savings.htm#indicator-chart

Cleaning

For each of our independent variables, we have the data from year 1960-2018 for 265 countries and regions. Here are improvements:

Improvement 1: Used only data from year 2010 - 2018:

a) Due to increasing globalisation and more countries relying on exports and imports of goods and services with time, the more recent the data the more similar economic activities the countries will have with each other.
b) Most of the missing data are between the years from early 1960s to 1990s.
c) A great recession occurred for 18 months between 2007 and 2009.

Improvement 2: Got rid of 'countries' that were not exactly countries, for control. E.g. Arab World, Caribbean small states, Central Europe and the Baltics, etc.

Improvement 3: Checked for identical data with various names or any data type issues.

Improvement 4: Dropped Male-to-population ratio as we had female-to-population ratio.

Improvement 5: Removed outliers for each independent variable, which we defined as data points that are more than 4 standard deviations from the mean.

Improvement 6: Dropped countries without savings rate for all 9 years. -> **Observation 1**

Improvement 7: Checked for correlation between the independent variables, shown by the correlation matrix in Fig1. For each pair with abs(correlation coeff) > 0.3, we eliminated one of the variables. We chose the variables to eliminate based on the number of available data points they provide. Dropped the following independent variables: unemployment, nominal_gdp, literacy_rate, female, education.

Observations of our data after cleaning:

Observation 1: Savings rates, had 518 total missing points. The following countries were dropped, represented in Fig 2 heatmap:

American Samoa, Andorra, Antigua and Barbuda, British Virgin Islands, Cayman Islands, Central African Republic, Chad, Cuba, Dominica, Equatorial Guinea, Equatorial Guinea, Fiji, French Polynesia, Gibraltar, Greenland, Grenada, Grenada, Iran, Islamic Rep., Isle of Man, Kiribati, Korea Dem. People's Rep., Libya, Liechtenstein, Maldives, Micronesia Fed. Sts., Monaco, Nauru, New Caledonia, Northern Mariana Islands, Papua New Guinea, Puerto Rico, Samoa, San Marino, Sao Tome and Principe, Sint Maarten (Dutch part), Solomon Islands, Somalia, St. Kitts and Nevis, St. Martin (French part), Syrian Arab Republic, Trinidad and Tobago, Turkmenistan, Turks and Caicos Islands, Tuvalu, United Arab Emirates, Virgin Islands (U.S.), Yemen Rep.

Observation 2: Literacy rate and Education have many missing values, globally, seen as column of reds in Fig 5.

Observation 3: Since countries without savings rate tended to also lack other input data, dropping data for those countries minimised data waste.

Observation 4: Finally we have 5 independent variables: age dependency ratio, GDP growth, nominal GDP per capita, population density, and total population.

Observation 5: We now have in total 1298 data points. The heat map in Fig 3. shows the number of data points belonging to each country.

Input Data Analysis

We concatenated all the data from 2010 to 2018, visualising data n a scatter graph for each independent variable with the savings rate. The **sole purpose is to check the data distribution** of the independent variables, **asserting the reliability** of it. We checked for outliers and unusual data points.

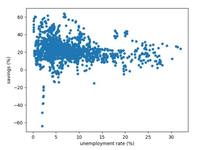
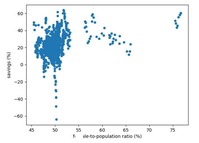
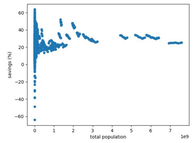


Fig 3. Scatter plot of savings rate with:

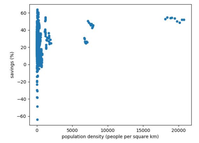
Unemployment rate
No obvious outliers. Data approximately ranges from 0 to 30%.



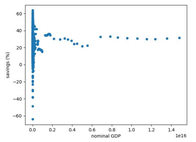
Male-to-population ratio
We observed some outliers here, with a country with 75+% male-to-population ratio. It is Qatar from 2010 to 2018. Oman, Maldives and Kuwait have 60+% male-to-population ratio in 2010 to 2018.
-> **Improvement 5**.



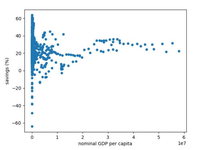
Total Population
Unusual data points with a total population larger than 2 billion (China at 1.3 billion). It was due to having "countries" that were not exactly countries, e.g. "World", "East Asia & Pacific", etc.
-> **Improvement 2**.



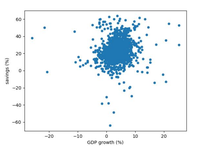
Population Density
We observed some outliers here, with Monaco and Macau SAR China with around 17000+ population density from 2010 to 2018. Singapore, Hong Kong SAR China in 2010 to 2018 had 5000+ population density.
-> **Improvement 5.**



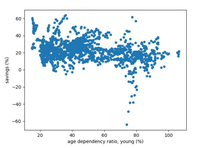
Nominal GDP
We observed some outliers here, with countries with 0.5x1e16+ nominal GDP (Indonesia and Iran). However, since the nominal GDP is in current local currency units relative to the U.S. dollar GDP, the difference in nominal GDP between countries does not matter. We are only seeing the correlation between the features and the savings rate.



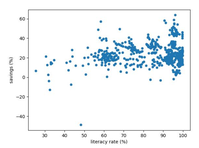
Nominal GDP per Capita
No obvious outliers. Data approximately ranges from 0 to 6x1e7 nominal GDP per capita. However, the distribution is very similar to Nominal GDP.
-> **Improvement 7**



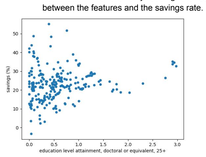
GDP growth
No obvious outliers. Data approximately ranges from -20 to 20%.



Age dependency ratio
No obvious outliers. Data approximately ranges from 10 to 106%.



Literacy rate
No obvious outliers. Data approximately ranges from 0 to 100%.



Education
No obvious outliers. Data approximately ranges from 0 to 3.

Fig 4. Scatter plot of each independent variable by savings rate, to check for data reliability

Fig 2. Savings Rates Missing Values per Country

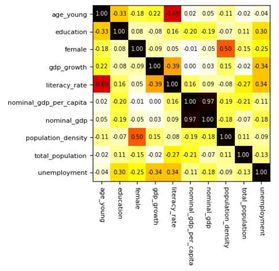


Fig 1. Correlation Matrix

Fig 3. Number of Available Data Points per Country

Other limitations:

- Analysis only done on countries. Hence the model is not indicative of anything geographically smaller or larger, e.g. individual households, continents.
- Few countries were dropped from the data. Check data cleaning for more information.
- Analysis based on a certain time frame. Hence the model is not indicative of anything outside of that time frame, say, effects on savings rate in few decades.

Statistics Analysis

We ran multiple linear regression, testing the effects of each variable, controlling for the other variables, with the savings rate. P value can tell us the probability of obtaining an effect equal to or more extreme than the one observed, assuming the null hypothesis is true, with significance level of $\alpha = 0.05$. Our hypothesis is as follows:

Null hypothesis:

There is **no** significant correlation between a <macroeconomic phenomenon> and the savings rate.

Alternative hypothesis:

There is a significant correlation between a <macroeconomic phenomenon> and the savings rate.

To do a multiple linear regression, we chose only the countries with all features present. This left us with only 90 data points out of 216 countries x 9 years = 1944 potential data points.. We were not able to identify any specific patterns as to which countries would have all features.

However, after **Improvement 6** with an aim to reduce potential noise, we ran multiple linear regression again, which consequently increased the number of useful data points. -> **Observation 4**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.410
Model:                            OLS   Adj. R-squared:                  0.327
Method:                 Least Squares   F-statistic:                     4.942
Date:                Sat, 25 Apr 2020   Prob (F-statistic):           4.84e-05
Time:                        15:11:20   Log-Likelihood:                -253.51
No. Observations:                  74   AIC:                             527.0
Df Residuals:                      64   BIC:                             550.1
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
age dependency ratio, young (%)  -0.2953      0.077     -3.845      0.000      -0.449      -0.142
education level attainment, doctoral or equivalent, 25+  -10.5282      6.518     -1.615      0.111     -23.550       2.494
female-to-population ratio (%)    0.9148      0.189      4.829      0.000       0.536       1.293
GDP growth (%)                     1.2377      0.438      2.846      0.006       0.369       2.106
literacy rate (%)                 -0.1570      0.080     -1.965      0.054      -0.317       0.003
nominal GDP per capita          5.966e-08   9.56e-07      0.062      0.950   -1.85e-06    1.97e-06
nominal GDP                     2.394e-15    1.4e-13      0.017      0.986   -2.78e-13    2.83e-13
population density (people per square km)    -0.0007      0.002     -0.293      0.770      -0.005       0.004
total population                2.15e-08   1.86e-08      1.154      0.253   -1.57e-08    5.87e-08
unemployment rate (%)             -0.0086      0.197     -0.501      0.618      -0.402       0.295
==============================================================================
Omnibus:                       13.870   Durbin-Watson:                   1.797
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               27.264
Skew:                           0.599   Prob(JB):                     1.20e-06
Kurtosis:                       5.729   Cond. No.                     3.34e+14
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.34e+14. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fig 6. Before **Improvement 6**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):           0.642
Model:                            OLS   Adj. R-squared (uncentered):      0.640
Method:                 Least Squares   F-statistic:                     369.8
Date:                Sat, 02 May 2020   Prob (F-statistic):          3.35e-227
Time:                        18:55:25   Log-Likelihood:                -4231.8
No. Observations:                1038   AIC:                             8474.
Df Residuals:                    1033   BIC:                             8498.
Df Model:                           5
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
age dependency ratio, young (%)  0.2070      0.014     14.557      0.000       0.179       0.235
GDP growth (%)                   1.3591      0.151      9.017      0.000       1.063       1.655
nominal GDP per capita        6.426e-08   1.52e-07      0.423      0.672   -2.33e-07    3.62e-07
population density (people per square km)    0.0046      0.001      6.702      0.000       0.003       0.006
total population              6.854e-08   9.98e-09      6.867      0.000     4.9e-08   8.81e-08
==============================================================================
Omnibus:                       32.414   Durbin-Watson:                   1.828
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               77.382
Skew:                          -0.062   Prob(JB):                     1.57e-17
Kurtosis:                       4.332   Cond. No.                     1.72e+07
==============================================================================
```

Fig 7. Final statistics analysis -> **Observation 3**

Conclusion

We **rejected** the null hypothesis for the following four independent variables.

Age dependency ratio of young working age, in %

Definition: "Age dependency ratio, young, is the ratio of younger dependents--people younger than 15--to the working-age population--those ages 15-64"

Test Statistics: p-value = 0.000, t-score = 14.557, correlation-coefficient = 0.2070

Explanation: With a correlation coefficient of 0.2070, this means that with a 1% increase in age dependency ratio, the savings rate is approximated to increase by 0.2070%. We think the result is reasonable, with reasons:
- Expectancy of having a child as an addition to the family, the working-age parents are likely to make plans to save a larger percentage of their income in order to support the additional spending
- Raising a young dependent often-times require expensive one-time expenditures on things such as education tuition, etc.

Limitations: Confounds that we have not controlled for. E.g. it might be the case that people make the decision of having children based on factors such as education and medical care qualities of their countries, which are then correlated with the cost of education and medical care, which then affect savings rates. In that case, we are not controlling for factors such as education cost when we in fact should be.

GDP growth, in %

Definition: "Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2010 U.S. dollars. "

Test Statistics: p-value = 0.000, t-score = 9.017, correlation-coefficient = 1.3591

Explanation: By the multiple linear regression, a 1% increase in the GDP growth will result in an increase in the savings rate by 1.3591%. The correlation coefficient of GDP growth is the highest among all factors put into the regression model. We think the result is reasonable, with reasons:
- People during 2010 to 2018 still suffered from the Great Depression, where higher GDP growth rate discouraged spending and encouraged people to save more. It could imply that people became more prepared, by saving up, in a likely event of the repeating economic crisis.

Limitations: Confounds that we have not controlled for, such as the impact from the 2008 financial crisis. For example, people in 2010 were still recovering from the crisis, so they did not have much money to save. People in 2018 were in a better shape, being able to save more portions of their income. The GDP growth rate slowly increased from 2010 to 2018 after the crisis was gone. In this case, the impact from the crisis is not controlled for, although it greatly influenced the savings rate. In addition, since the real world does not include only the post-period of an economic crisis, the correlation can change with different states of the world.

Total population

Definition: "Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are mid-year estimates."

Test Statistics: p-value = 0.000, t-score = 6.868, correlation-coefficient = 6.854e-08

Explanation: The correlation coefficient for total population is 6.854 x 10^-8 (or 0.00000006854). In other words, with total population increasing by 1, savings rate increases by 0.00000006854%. Coefficient is relatively small due to total population contains several digits, while savings rate is at maximum 100. We think the result is reasonable, with reasons:
- The large number of people working-age population has to feed. With more mouths to feed in a household, working-age people have to be more considerate on spending, leading to higher savings rate.
- More people in a household implies that there are shared consumptions, such as furnitures, so spending per person will be lower, driving savings rate up.

Limitations: Confounds that we have not controlled for. For example, total population can depend on the medical advancement. With greater medical technologies comes more lives saved and higher expenses in attending alive people, affecting savings rate. We have not controlled for possible confounds like medical advancement.

Population density

Definition (as given by the World Bank): "Population density is midyear population divided by land area in square kilometers. "

Test Statistics: p-value = 0.000, t-score = 6.702, correlation-coefficient = 0.0046

Explanation: As for population density, the correlation coefficient is 0.0046, meaning that with one more person added to each square kilometer of land area, savings rate increases by 0.0046%. We think the result is reasonable, with reasons:
- Cost of living that comes from population density. In densely populated areas, there is more demand for consumptions, such as food, medical care, and other necessities. Since supply is relatively limited, the cost of living is higher in these areas. Residents of such areas adjust themselves to save up more to combat a high living cost. In an area with low population density, saving is not as much needed due to the low cost of living. Thus, residents in these areas have lower savings rates.

Limitations: Possible confounds we have not controlled for, such as habits of people living in different population density areas. With more career opportunities in an area, people save money to move or commute to that area in order to pursue their career interest, for example. That area becomes more densely populated as a result, while the savings rate is also higher, due to the type of people living there. In such cases, we have not controlled for possible confounds.

We **failed to reject** the null hypothesis for the following independent variable.

Nominal GDP per Capita

Definition: "GDP per capita based on purchasing power parity (PPP)."

Test Statistics: p-value = 0.672, t-score = 0.423, correlation-coefficient = 6.426e-08

Explanation: We failed to reject the null hypothesis because the p-value 0.672 is greater than our significance level 0.05.

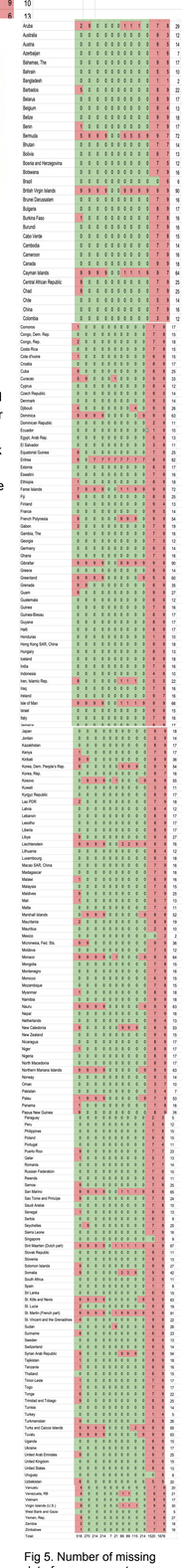


Fig 5. Number of missing data for every country, per feature. -> **Observation 2**, **Observation 3**