# The Use of Socioeconomic Indicators to Estimate PPP Adjusted Income: An Application to Cuba Circa 1957

by

Luis Locay
University of Miami
mambises@bellsouth.net

December 2021

Preliminary Draft

I wish to thank Ethan Gropper for able research assistance.

# The Use of Socioeconomic Indicators to Estimate PPP Adjusted Income: An Application to Cuba Circa 1957

It is not uncommon to encounter a situation where internationally comparable (purchasing power parity [PPP] adjusted) measures of per capita income are questionable, or even missing, for a country for which other socioeconomic variables are available. The purpose of this paper is to develop techniques using information from countries that have data on both PPP adjusted income and other socioeconomic variables to estimate PPP adjusted per capita income for a country that is missing the former. These techniques are applied to Cuba in 1957.

Interest in Cuba has often exceeded what one would expect based solely on its small size. This has been especially so since the Castro Revolution of 1959. A major topic of interest surrounding Cuba has been the comparison of pre- and post-revolutionary economic performance, especially income per capita. Any pre- and post-revolutionary comparisons require, of course, not only recent PPP adjusted income - on which most recent discussions focus - but also PPP adjusted income from the late pre-Revolutionary period.

Until Ward and Devereux (2012) (henceforth WD) we did not have a direct estimate of pre-Revolutionary PPP adjusted income for Cuba. While their work is grounded in sound methodology, problems with missing data still require them to make a variety of assumptions. More recently the Maddison Project Database (henceforth Maddison) provides an income measure for Cuba, though their methodology is not known to me. The two measures are substantially different. WD give a value for per capita GDP that is 27% that of the US in 1955. The 27%, however, is most likely a lower bound, as on all occasions where they have to make assumptions, they make conservative ones. They estimate consumption to have been 35% that of the US. Maddison (I use the 2020 database) gives an income per capita for Cuba in 1955 that is only 15% that of the US. One of the primary aims of this paper will be to see if the indirect method developed here using available data on other socioeconomic variables supports either of these estimates.

The year 1957 was chosen as the last pre-revolutionary year before the Revolution is likely to have had significant economic effects. This will not substantially matter for the WD vs Maddison comparison.

The paper is organized as follows. The next section briefly discusses the data and describes Cuba's relative position for several of the socioeconomic indicators. Section 2 introduces and discusses the statistical model, and section 3 presents the formal results. I end with a discussion of the results and future work in section 4.

## 1. The Data

The dataset consists of 118 countries, including Cuba, for which Maddison provides PPP measures of relative per capita income in 1957, and for which I was able to find or construct data on 11 other social, economic, or demographic variables.[1] Centrally planned economies were also excluded. The variables are described in Table 1, along with their corresponding values for the dataset, Cuba, the US, and some comparison countries from the region of the Caribbean. Also included in Table 1 are three variables that were not included in the analysis at this time due to missing values. These are televisions, newspaper circulation, and schooling levels. The sources of the data are various UN Statistical Yearbooks, 1957-1978, the World Bank Databank, Barro and Lee Educational

[1] The exception was Kuwait. It was excluded because Maddison reports an extremely high level of income.

Attainment Dataset, and Abouharb and Kimball (2007).[2] I will provide a detailed description of sources and construction of the dataset in the future.

According to Maddison, Cuban GDP per capita in 1957 was 76% of Costa Rica's, 25% of Venezuela's, and only 17% of the US level. It was 41% higher than the Dominican Republic's. Even a casual inspection of Table 1 shows that at least for the Caribbean countries the social economic indices are not supportive of Maddison's relative income rankings. Cuba outperforms Costa Rica and the DR on every single measure, sometimes by large amounts. It even outperforms Venezuela on all measures but telephones, cars, and electricity consumption, and in all but the last one the two countries are close.

Figures 1 – 4 illustrates Cuba's standing among the individual countries in Table 1 and the rest of the countries in the dataset on four of the socioeconomic indices – telephones, automobiles, female life expectancy, and infant mortality. The straight line in black is always the regression of the log of the index on the log of GDP/capita (according to Maddison), computed on the entire dataset except Cub. Cuba is situated on the regression line according to its value for the vertical axis. The corresponding horizontal axis value is the log of GDP/capita that its value for each index implies. For example, in 1957 Cuba had 40.3 telephone lines per 1,000 persons over 18, or 6.36 in logs. The log of GDP per capita that would correspond to 6.36 from the regression line is 8.54, or a GDP per capita of $5,101. This is 29% of Maddison's GDP per capita in 1957 for the US of $17,406.

[2] For Cuban infant mortality I relied on estimates in McGuire and Frankel (2005).

The implied Cuban GDP per capita from Figures 2 - 4 for cars, female life expectancy and infant mortality rate are \$7,298 (42% of that of the US), \$7,476 (43% of that of the US), and \$12,690 (73% of that of the US), respectively. There is nothing in Table 1 of Figures 1 - 4 to suggest that in 1957 Cuba was poorer than Costa Rica, much poorer than Venezuela, and had a GDP per capita only 17% of the US level. The formal analysis will support this.

## 2. The Statistical Model

Let $z_i$ be the level of some private or public good in a given country. A typical simple aggregate demand function would relate $z_i$ to average income, $y$, relevant prices, $p_i$, and country specific factors, $x$. Suppose there are *n* such goods. A log linear approximation, where $z_i, y$ and $p_i$ are in logs, would give rise to the following set of equations:

$$z_i = \alpha_i x + \lambda y + \delta p_i + \varepsilon_i,\ i = 1,...,n \tag{1}$$

Since I have no data on prices, I assume that prices are subsumed in the error term. Doing so gives rise to a new system of *n* equations:

$$z_i = \beta_i x + \theta y + u_i,\ i = 1,...,n \tag{2}$$

Notice that the coefficients on the remaining variables have been renamed. This is because prices may not be independent of the other variables in the equations. This means that (2) should not be interpreted as structural equations.

Let there be $j = 1,...,m$ countries. Suppose that for country $\ell$ we do not observe income. Let $I_\ell$ be an indicator function that takes the value one for country $\ell$, and zero otherwise. We can rewrite (2) as follows:

$$z_{ij} = \beta_i x_j + \theta\left(I_\ell \psi + (1 - I_\ell) y_j\right) + u_i,\ i = 1, \ldots, n,\ \ j = 1, \ldots, m \qquad (3)$$

Where $\psi$ is the unobserved income of country $\ell$ and is treated as a parameter to be estimated. We can estimate (3) as a set of non-linear seemingly unrelated equations. This method will give us an estimate of the log of income for country $\ell$, as well as confidence intervals for that estimate.

The procedure described above for country $\ell$ can be repeated successively for each country in the dataset. In each iteration the chosen country is treated as if its income is missing. The predicted and actual log incomes can then be compared to each other, and the distribution of errors can be used to construct other estimate and confidence intervals.

## 3. The Formal Results

Before discussing the findings, a comment on the elements of *x*, the country specific factors. At this point I have only included, besides a constant, an indicator variable that takes the value one if I deemed the country to have been significantly and negatively impacted by major conflicts around the time of WW II.[3] My reasoning was as follows: The levels of the socioeconomic indicators, *z*, are probably more related to permanent than current income, and countries severely impacted by war 10-15 years prior to 1957 would be further below their long-run income path than those that had not. I'll say a bit more about this in the conclusion.

Table 2 summarizes some of the results of the estimation of (3) by non-linear seemingly unrelated least squares. The column labelled "Income Elasticity" corresponds to parameter $\theta$ in (3). It measures the impact of the log of GDP per capita on the

[3] Besides WWII, I included the Spanish Civil War and the Algerian War of Independence.

corresponding log of the socioeconomic index listed. All the $\theta$ parameters are of the expected sign and highly significant.

The point estimate of the log Cuban GDP per capita (not shown) is 8.92, which corresponds to a GDP per capita of $7,461, or 43% of that of the US. This is significantly higher than the WD estimate of 27%, not to mention that of Maddison at 17%.[4] In levels, the 95% confidence interval is (3,724, 14,947), or 21% - 86% the US level. The WD estimate of 27% of the US level is within the 95% confidence level, but Maddison's 17% is not.

I do not know if the procedure used here results in an unbiased estimate of the log of Cuban GDP per capita, or the error terms of $\hat{\psi} = 8.92$ are distributed according to Student's t-distribution. To explore the distribution of this parameter estimate, the same procedure was used for every country in the dataset under the assumption that its per capita GDP was unobserved.[5]

The sample consisted of 117 countries. The average log of GDP per capita (from Maddison) is 7.924. The average of the estimates was almost identical at 7.926.[6] Using the actual distribution of the estimated error terms, the 95% confidence interval is (3507, 18,535). Again, the Maddison value is outside of the interval, but WD is not. Another way to look at is to ask if the Maddison number is correct, what is the probability of getting an estimation error equal to or greater than the one obtained here? That probability is 0.5%. For the WD GDP estimate it is 13% and for their consumption estimate it is 27%.

[4] It isn't so much higher than WD's estimate for consumption at 35%.

[5] Cuba was excluded from the dataset for these calculations.

[6] The averages of the levels are different. The average of the data is $4,184, while the average of the estimates is $4,825.

## 4. Conclusion

Cuba's relative international socioeconomic standing in 1957 as reflected in various socioeconomic variables is not consistent with the low GDP per capita reported in Maddison. It is more consistent with the estimate provided by WD, though even theirs appears to be too low. There are several reasons, however, why the WD estimate may more consistent with that obtained in this paper than it at first appears.

First, all the relative GDP data comes from Maddison. The procedure used in this paper is really a sort of consistency check on a set of income values. What we have found is that Maddison's reported GDP per capita figures for Cuba and the US are not consistent with other observed socioeconomic variables. A more appropriate comparison would involve applying the same procedure to the WD dataset. This is left for future work, as well as an application to a third dataset that does not provide information for Cuba – The Penn World Tables.

Second, the socioeconomic indices should be more related to measures of income than to GDP. These two tend to be similar, but in Cuba's case they may have differed more than usual. Furthermore, permanent income – or what amounts to the same thing, consumption – is likely a better measure to use in the estimation than current income (or current GDP), which would help explain why the WD value for relative consumption is closer to the estimate of this paper.

Third, WD purposely made assumptions in their estimation that they considered "conservative". That is assumptions that would tend to underestimate Cuban relative income and consumption.

The work presented here excludes several socioeconomic indices, mostly because using them would involve dropping several countries from the analysis. There seems to be several ways to proceed. One way would be to try to construct estimates of the missing values. I believe this approach is not worth the effort, especially in light of the next two alternatives. A second approach is to treat the set of equations as an unbalanced panel, so that not all countries are necessarily represented for each socioeconomic index. Originally, I wanted to include as many countries as possible to get a complete picture of the relationship between income and the socioeconomic indices at all levels. While this is valuable in the estimation of a structural model, that is not what is being done here. The estimates of income may be more accurate if we limit the analysis to countries that are roughly in the same range as Cuba's. Afterall, the log linear approximation is probably better over a narrower range of incomes. This would exclude the poorest countries, which are the most likely to have missing or poor data.

Finally, there are the country specific characteristics that may improve the estimation. Some geographic or climate variables come to mind. Another consideration is how the WW II variable appears in the model. Currently it enters each equation independently. This may not be the best way given the interpretation the variable was given.

## References

Abouharb, M. R. and A. L. Kimball (2007). "A New Dataset on Infant Mortality Rates, 1816–2002," *Journal of Peace Research*, vol. 44, no. 6, 2007, pp. 745–756.

Barro and Lee Educational Attainment Dataset.

McGuire, James W. and Frankel, Laura B. "Mortality Decline in Cuba, 1900-1959: Patterns, Comparisons, and Causes." *Latin American Research Review*, vol. 40, no. 2 (2005): 83-116.

UN Statistical Yearbooks, 1957-1978.

Ward, M. and J. Devereux **(**2012). “The Road not taken: Pre-Revolutionary Cuban Living Standards in Comparative Perspective,” *The Journal of Economic History*, vol. 72, no. 1, pp. 104-133.

World Bank Databank.

# Tables

Table 1 - Socioeconomic Variables

| Name | Description | Average | Cuba | US | Costa Rica | DR | Venezuela |
|---|---|---|---|---|---|---|---|
| Phones | Telephone lines per 1000 persons 18 and over. | 57.7 | 40.3 | 576.6 | 22.5 | 11.1 | 44.4 |
| Radios | Radios per 1000 persons. | 81.7 | 165.6 | 876.2 | 67.4 | 34.2 | 111.9 |
| LifeEF | Female life expectancy at birth. | 53.1 | 64.2 | 73 | 61.5 | 51.4 | 59.6 |
| LifeEM | Male life expectancy at birth. | 49.9 | 60.8 | 66.6 | 58.8 | 48.6 | 56.6 |
| IMR | Infant mortality rate. | 119.8 | 37 | 27.3 | 51.1 | 138.7 | 91.1 |
| Doctors | Physicians per 1000 persons. | 0.41 | 0.97 | 1.25 | 0.34 | 0.17 | 0.55 |
| Dentists | Dentists per 10,000 persons. | 0.15 | 0.32 | 0.53 | 0.10 | 0.02 | 0.09 |
| TFR | Total fertility rate. | 5.64 | 3.7 | 3.67 | 7.11 | 7.64 | 6.46 |
| Cars18 | Automobiles per 1000 persons 18 and over. | 35.1 | 41.9 | 505.1 | 23.3 | 7.0 | 59.1 |
| Electricity | Metric tons of coal equivalent consumed per person. | 0.8 | 0.71 | 7.8 | 0.22 | 0.15 | 2.37 |
| Kcal | Kcal per person per day. | 2351 | 2740 | 3100 | 2406 | 2062 | 2088 |
| TV | Televisions per 1000 persons. |  | 45.2 | 274.6 | 2.1 | 4.5 | 15.7 |
| Newsp | Newspaper circulation per 1000 persons over 18 |  | 215.7 | 522.3 | 173.2 | 56.1 | 172.2 |
| Schooling | Average years of total schooling. |  | 3.9 | 8.9 | 3.9 | 2.7 | 2.8 |

Table -2

| Dependent Variable in Logs | Income Elasticity | t-value | R2 |
|---:|---:|---:|---:|
| Phones | 1.87 | 23.2 | 0.83 |
| Radios | 1.88 | 19.4 | 0.78 |
| LifeEF | 0.22 | 14.3 | 0.67 |
| LifeEM | 0.21 | 14.0 | 0.66 |
| IMR | -0.62 | 14.0 | 0.65 |
| Doctors | 1.39 | 17.5 | 0.74 |
| Dentists | 2.14 | 18.2 | 0.75 |
| TFR | -0.27 | 11.0 | 0.58 |
| Cars18 | 1.5 | 14.7 | 0.65 |
| Electricity | 1.8 | 19.9 | 0.78 |
| Kcal | 0.15 | 10.8 | 0.52 |

Estimated Cuban GDP/Capita

| Log GDP/Capita | GDP/Capita | Standard Error |
|:---:|:---:|:---:|
| 8.917 | 7,461 | 0.355 |

# Figures

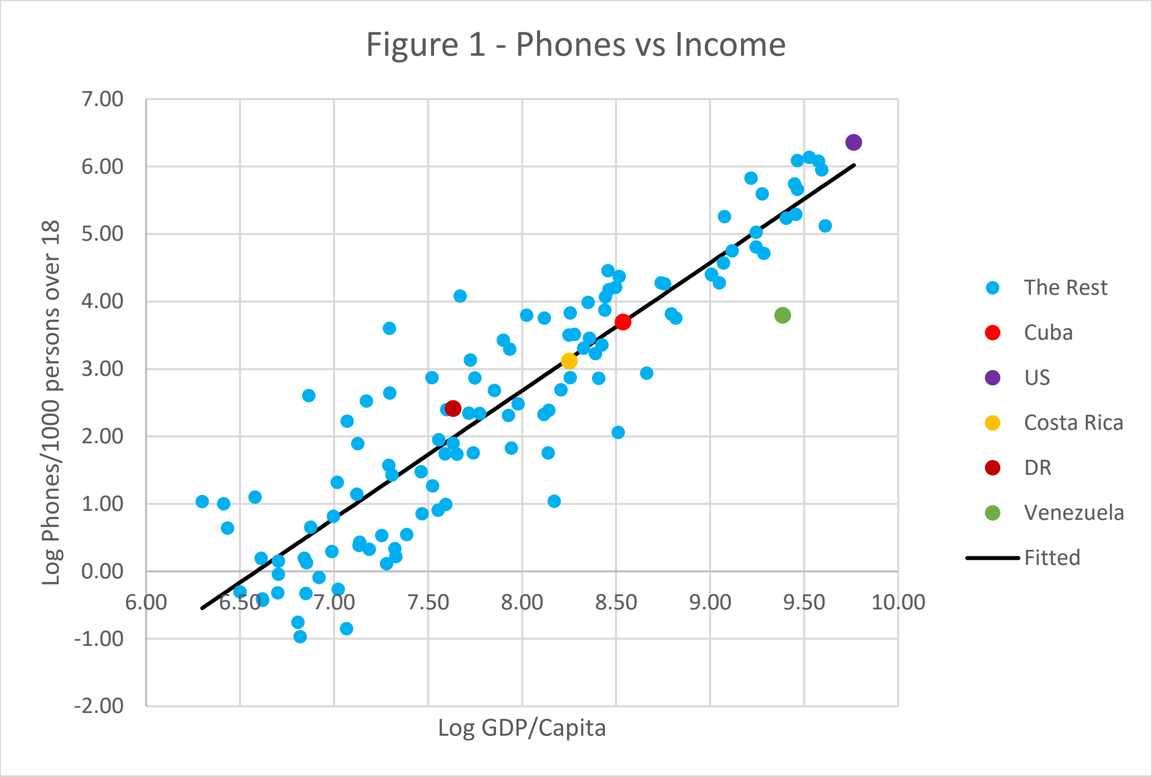
Figure 1 - Phones vs Income
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
6.00
6.50
7.00
7.50
8.00
8.50
9.00
9.50
10.00
Log Phones/1000 persons over 18
Log GDP/Capita
The Rest
Cuba
US
Costa Rica
DR
Venezuela
Fitted

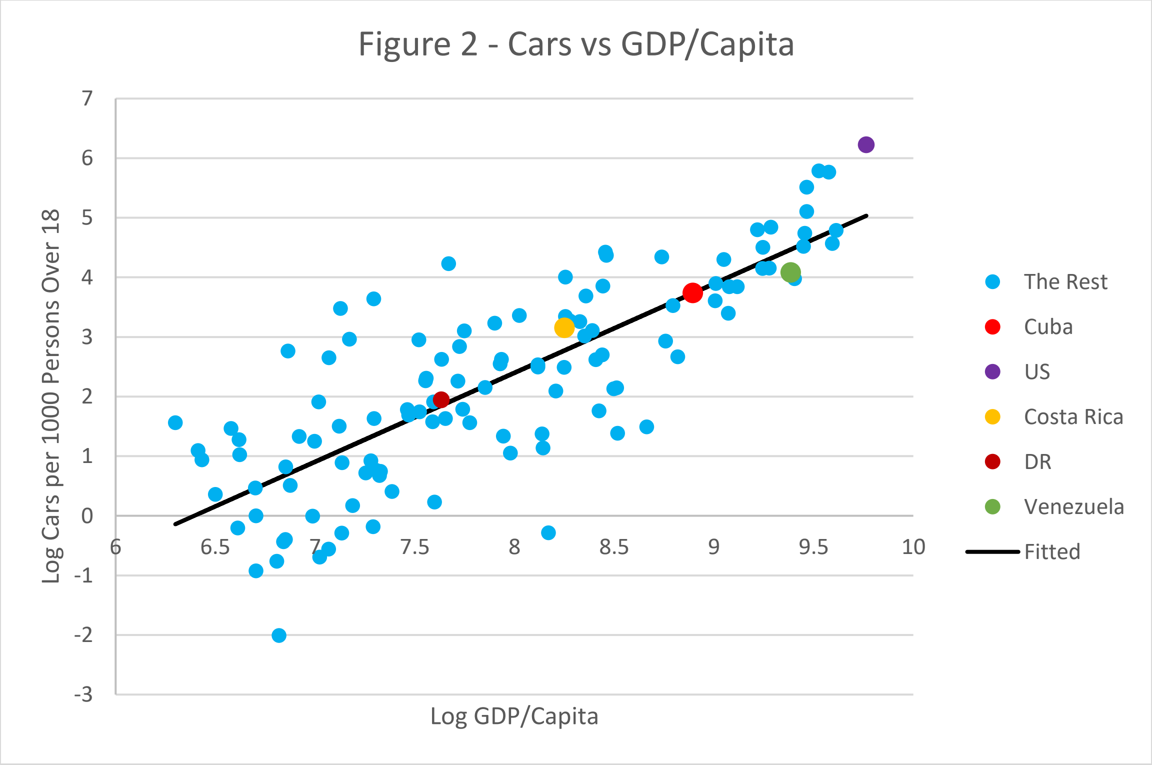
Figure 2 - Cars vs GDP/Capita
7
6
5
4
3
2
1
0
-1
-2
-3
6
6.5
7
7.5
8
8.5
9
9.5
10
Log Cars per 1000 Persons Over 18
Log GDP/Capita
The Rest
Cuba
US
Costa Rica
DR
Venezuela
Fitted

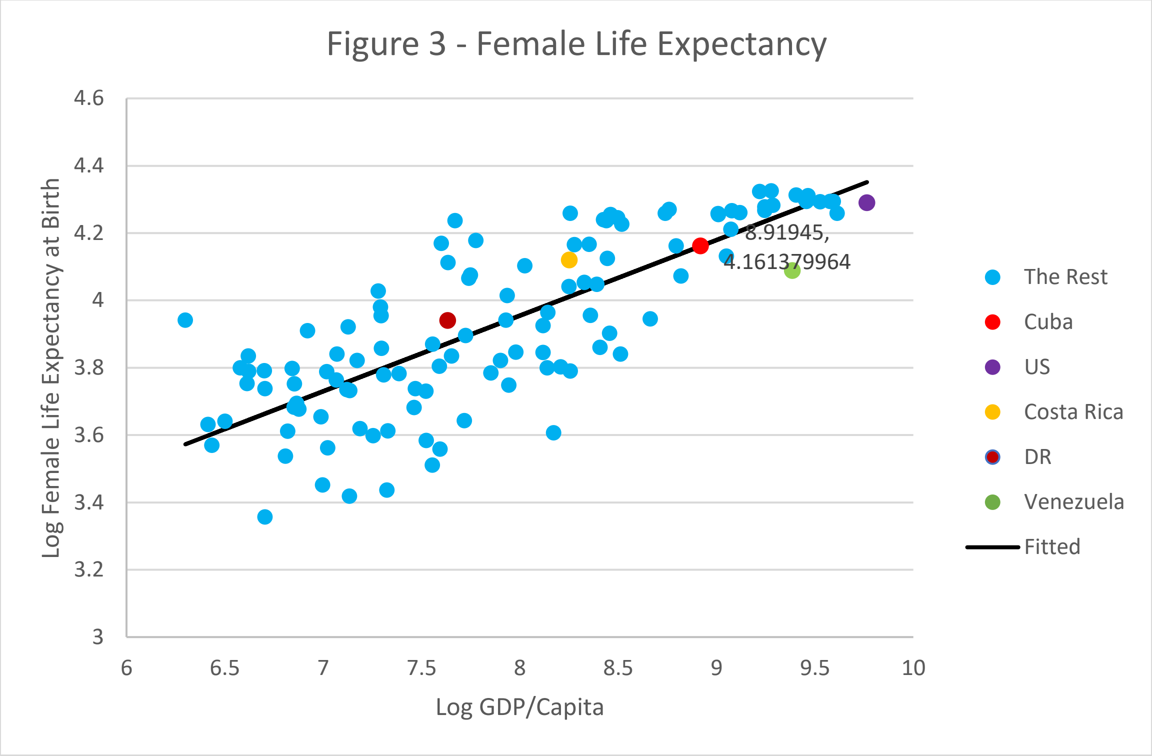
Figure 3 - Female Life Expectancy
Log Female Life Expectancy at Birth
4.6
4.4
4.2
4
3.8
3.6
3.4
3.2
3
6
6.5
7
7.5
8
8.5
9
9.5
10
Log GDP/Capita
8.91945,
4.161379964
The Rest
Cuba
US
Costa Rica
DR
Venezuela
Fitted

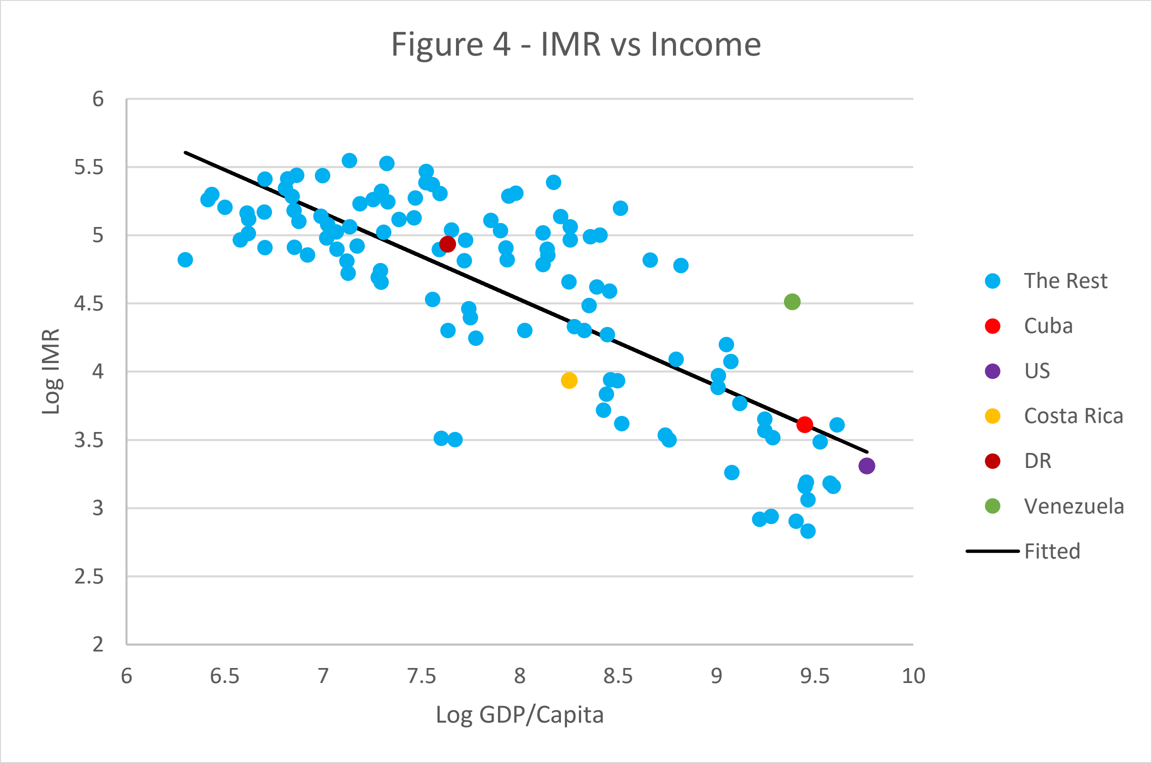
Figure 4 - IMR vs Income
Log IMR
6
5.5
5
4.5
4
3.5
3
2.5
2
6
6.5
7
7.5
8
8.5
9
9.5
10
Log GDP/Capita
The Rest
Cuba
US
Costa Rica
DR
Venezuela
Fitted