The Relationship Between Modern Literacy Levels and Income Per Capita Based on Proximity to Jesuit Mission Sites*

Arshnoor Kaur Gill

09/12/2020

Abstract

In this paper we attempt to reproduce 'Human Capital Transmission, Economic Persistance and Culture in South America' by Felipe Valencia Caicedo. We found that though there is a 0.4854% increase in modern literacy levels and a 1.249e-01 log increase in the income per capita if one lives in a municipality 100 km closer to a Jesuit mission site in Brazil, Paraguay and Argentina, the OLS regression considers this variable insignificant. This suggests other variables included in the regression, such as access to rivers, may influence these measures of human capital more so than distance from a mission site. A weakness is that through using Jesuit mission sites as proxies for sources of investment in this community, a false linkage could be created between the presence of missionaries and economic and social development in a municipality.

Keywords: Modern literacy, linear regression, South America, income per capita

1 Introduction

The activities of missionaries in developing countries is a hotly contested issue in our modern context. Though all religions have delegates who attempt to spread their faith, this is most associated with the Christian religion, especially in times of colonial expansion. Some believe that the work of building physical infastructure such as schools and hospitals that missionaries associate themselves with justifies the presence of these religious representatives in developing nations. However, there is the other perspective that missionary work is simply a modern by-product of imperialism, which ultimately results in hurting Indigenous populations through unintended consequences of contact, such as contracted disease. Some countries like Nepal have in fact labelled this practice illegal (Luckhurst). What this reproduction of this paper intends to do is examine the effect of Jesuit missions in Argentina, Brazil and Paraguay, in a set of missions called the "reductions" before the Jesuits were then ordered out of South America by the then king of Spain, in order to examine the longlasting effects on the local populations.

The way these findings were derived was through the use of a standard linear regression, the distance from Jesuit mission sites in kiometers used as the explanatory variable to attempt to predict (a) the 2000-2002 literacy rate as a percentage of the municipality (15 or 25 and older), and (b) the ln of the income per capita in 2000 of a municipality. The data is from the initial report published by Valencia Caicedo, which from him was collected from several sourses including: the national censes from Brazil, Paraguay and Argintina, as well as Jesuit Archives and firsthand correspondances from Jesuit officials in Rome (Valencia Caicedo). There is a 0.4854% increase in modern literacy levels and a 1.249e-01 log increase in the income per capita if one lives in a municipality 100 km closer to a Jesuit mission site in Brazil, Paraguay and Argentina, but this is insignificant in terms of the p-values fo the individual regression coefficients, contrasting the findings of the initial report.

*Code and data are available at https://github.com/arshnoor123/reproduction-of-caicedo.

The outline of the paper is as follows. In the **Data** section, I further with reference to Valencia Caicedo's paper discuss the methodology behind the data collection utilized in this report. Within this section, I deliberate on the strengths and purposes of the data for the purpose of this regression analysis, as well as touch on summary statistics for variables of interests. In the following section, **Model**, I lay out the regression equations that I will be utilize for the OLS analysis. In the **Results** section, there will be graphical and tabulated representation of the three relationships being investigated, and in the **Discussion** section the implications and legitimacy of the results in terms of significance will be deliberated, along with strengths and weaknesses of this reproduction. The most inherent of the weaknesses of this reproduction is that there may be a missing cofounder that creates a false relationship between the presence of Jesuit mission sites and literacy levels or economic prosperity of a location, as well as the lack of data regarding income effects in Argentina. For instance, it may not be anything in particular regarding the Jesuit missions but rather just any investment regardless of source in a community in terms of external aide or money that may lead to eventual relative prosperity. However, on the other hand, a key strength of the data is innovatively simplifying complex geographical calculations into a specific numerical variable, the difference from a mission site, alowing for straight-forward and easily interpretable analysis.

2 Data

This data, initially collected by Caicedo, comes from a variety of different sources. Covering Misiones and Corrientes within Argentina, Misiones and Itapúa within Paraguay and Rio Grande du Sol within Brazil, where the Jesuit mission sites were (based on historical analysis), the 578 observations are all regarding the community on a municipal level.

The data Caicedo uses regarding literacy rates (the percentage of the population of a municipality which is literate) and income (in terms of the Brazilian currency) comes from the national censes of these countries. For Brazil, this is the Brazilian Institute of Geography and Statistics (IBEG), the agency responsible for undertaking the census in Brazil. It appears that there is a discrepency for the literacy rate, in that for some municipalities it may have been collected for people fifteen or older, and other municipalities collected for people twenty-five or older. This sampling technique is described as "probabilistic", with the frame being permanent residents who live in private homes and the population being every citizen in Brazil. The technique appears to be systemic in intent, with the chosen housholds rotated on a yearly basis for whom will be picked for the sample ("Continuous National Household Sample Survey - Continuous PNAD"). For Paraguay, the literacy data comes from the Dirección General de Estadìstica, Encuestas y Censos (DGEEC), the national statistical agency, for the population per municipality 10 years or older in 2002. This appears to be probabilistic as well, but I was unable to recover the exact sampling technique undergone for the collection of this data. In Argentina, the literacy rate is from the National Institute of Statistics and Census of Argentina, for the population (10 years or older), in 2001. This data was collected through sampling individuals, households and dwellings ("OECD Assessment of the Statistical System of Argentina and Key Statistics of Argentina") as the frame, with the total population of Argentina as the population. There are obvious weaknesses in accumulating this data in this way—not only do they come from different agencies from different years, the cut-offs for the literacy are also in fact different. That being said, the data is from a time relatively close together, and the analysis later may be able to account for state effects.

In terms of income, direct income data could only be derived for Brazil, from the Institute of Applied Economic Research for 2000, in terms of annual earnings in the domestic currency for 2000, with the same sampling technique as described earlier. For Paraguay, data was collected through the World Bank for 2008 in the same technique and scope such that it could be compared to the Brazil observations on a muncipal level. That being said, there was no data in terms of income available for Argentinian municipalities in this time period, so the data regarding income in terms of distance from Jesuit sites will ignore Argentina. Income as a variable was transformed to "ln" of income, the natural logarithm, both because it makes the relationship with income further linear, and also because it makes much more intuitive sense to consider how the distance from a Jesuit site would change income in terms of percentage rather than dollars. There are evident weaknesses to this—through a lack of information regarding income for Argentinian municipalities, we cannot account for the direct impact that the sites had in an area where they were highly prominent.

However, a strength is the conversion to the ln of income—not only does it follow the accepted standard when it comes to economic analysis, using the natural logarithm also just makes sense. Someone who is making 50,000 dollars a year would be much more impacted by a wage increase of 1,000 dollars, versus someone making 500,000 dollars a year. The quality of life is not impacted uniformly by a unit increase of income regardless of where you are on the income scale, so using percentages universalizes the experience somewhat. Regardless of how much money you make, it makes intuitive sense that a 10% increase would be of same relative importance to you as someone who makes half as much or double your yearly income.

The variables of important for the purpose of this analysis are: the distance from a Jesuit site each municipality is from (in kilometres), the literacy rate of each municipality, from the year between 2000 and 2002 that it was collected, and the ln of the yearly income in the Brazilian currency. For literacy, there were a wealth of options of how this could've been measured—for instance, one variable from Caicedo was centered around "CFE"s, or country-fixed effects. These effects essentially take into account dummy variables for the country that the municipality is a part of, and uses it to affect the literacy variable. I've decided that I will use the literacy rates without the CFEs, such that it is less transformed and more "raw" data, with the intention to use the country each municipality is from as a categorical variable in the linear regressions that will be undergone.

Moreover, other variables that will be considered in the analysis are variables which represent geographic information, including: the country, the distance from a river, the precipitation levels and the distance from the coast. This geographical data was derived from the work of the organization BIOCLIM. The intent is hopefully to use these variables to stabalize any effect terrain may have had to access to information.

Below is distribution of the income of the observations from Brazil and Paraguay:

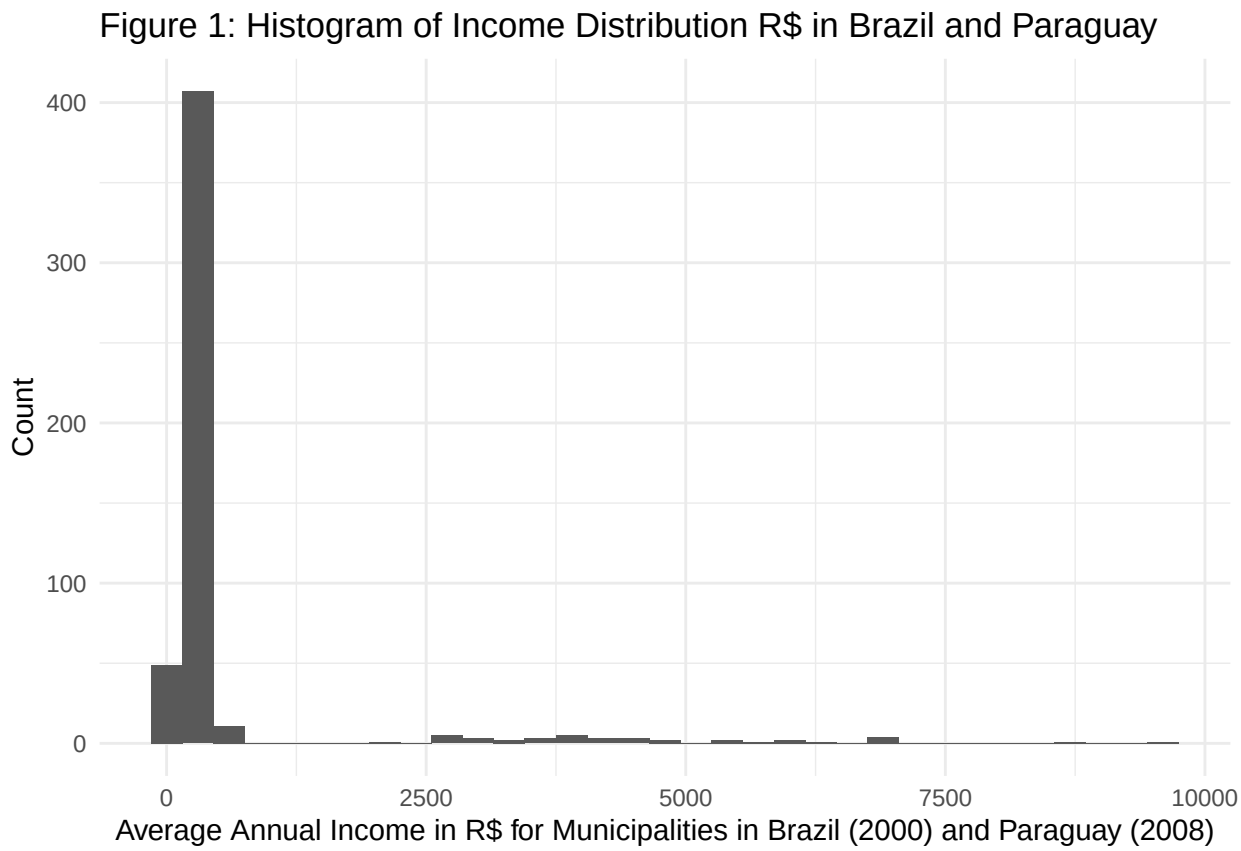


Figure 1: Histogram of Income

It is encouraging that income (not the ln of the income) appears to, with small outliers, follow a relatively normal distribution, heavily right-skewed. The fact that a predictor variable behaves in this fashion indicades a regression analysis may be appropriate. Moreover, it is worth noting the fact that much of the observations that have been collected are in areas with a very low average income, so this begs the question of whether missionaries flocked to areas where people were poorer and therefore more in need of aid.

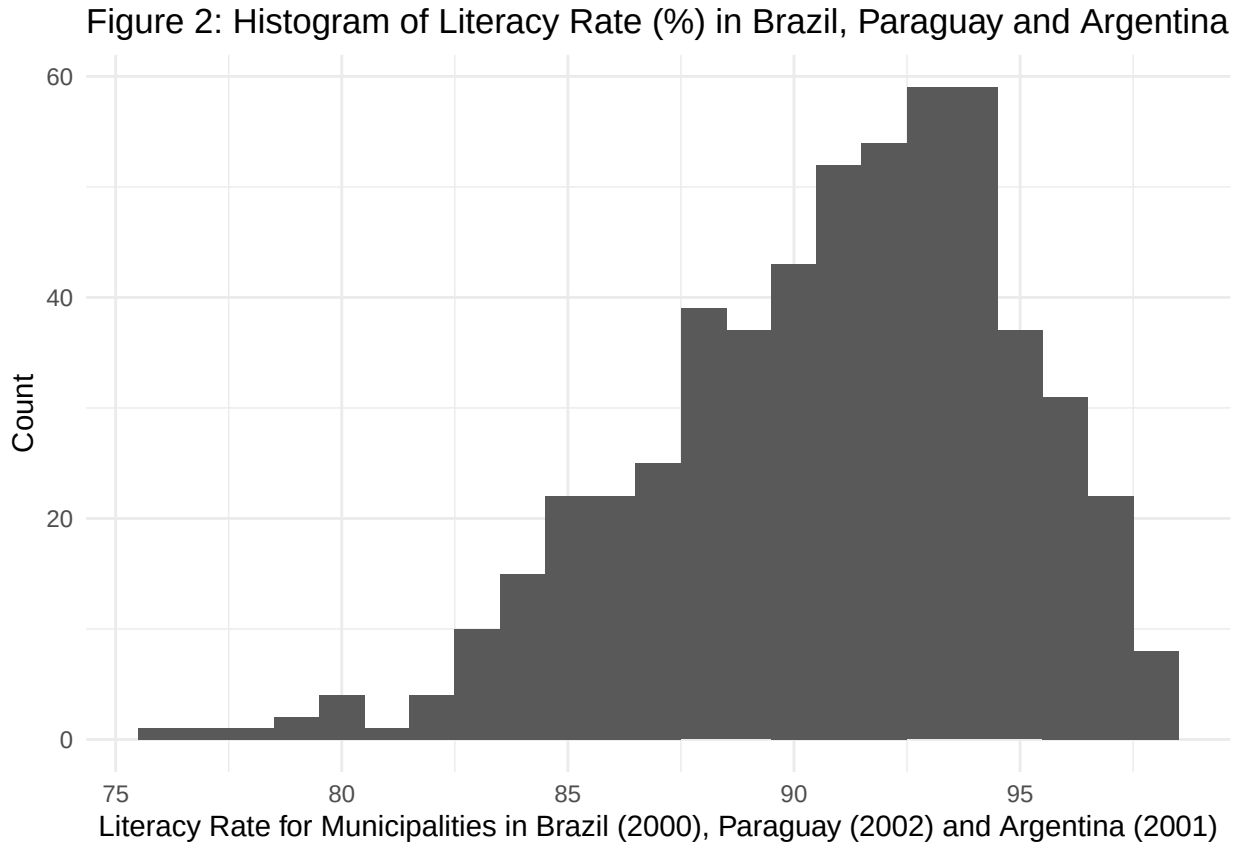


Figure 2: Histogram of Modern Literacy Rates

Here is a histogram of the other response variable we'll be looking at, the modern literacy rate in Brazil and Paraguay. As you can see, this also follows a relatively normal distribution with a left skew, the peak at a literacy rate greater than 92.5%. This indicates that the population we'll be investigating through OLS regression on the sample is broadly educated and that perhaps the effect the distance from a Jesuit mission has on the literacy rate will be lower than anticipated given how broadly the results appear to colesce around 90 to 95% for the overall dataset.

Finally, to round out the analysis of the response variables, below is a table of summary statistics in order to examine the spread and centre of the data. For the variable income, the big difference between the average income and the median income is stark, indicating that huge outliers are skewing the mean such that it is not truly representative of the broad majority of the sample. In contrast, for literacy rates, the mean and median appear about the same, which reflects what we saw in the histogram of a overall normal distribution.

Table 1: Summary Statistics of Response Variables

	Income (R$)	Literacy Rate (%)
Mean	5.789704e+02	90.83551
Minimum	8.479019e+01	75.67600

25th Quantile	1.880417e+02	87.97100
Median	2.441659e+02	91.49400
75th Quantile	3.114556e+02	93.95950
Maximum	9.637997e+03	98.40500
Variation	1.566601e+06	17.66385

3 Model

In order to graph the relations between the predictor variable, distance from a Jesuit mission, and the two response variables described thus far, (a) the literacy rate and (b) natural logarithm of the income of the sampled municipalities. Both relations will be derived through the use of ordinary least squares regression analysis, using the *lm()* function. This is a frequentist approach and relies on the concept of fixed parameters in the following equations:

Firstly,

$$\text{literacy rate} = \beta_0 + \beta_1 \cdot \text{distance from Jesuit mission} + \beta_2 \cdot \text{Brazil} + \beta_3 \cdot \text{Paraguay} + \beta_4 \cdot \text{distance from river}$$

$$+\beta_5 \cdot \text{distance from coast} + \beta_6 \cdot \text{precipitation level} + \beta_7 \cdot \text{altitude} + \text{error}$$

In which:

- `Beta 0` is the intercept parameter, in which the distance from a Jesuit mission is 0 km.
- `Beta 1` is the slope parameter, which represents how the percentage of the total population's literacy would change with a one kilometer difference in km.
- `distance from Jesuit mission` is thus the km the municipality is from a Jesuit mission.
- `Beta 2` is the difference between the expected value of the literacy at any level if the observation is from Brazil versus not from Brazil.
- `Brazil` is a dummy variable which equals "0" if the municipality being observed is in not in Brazil, and "1" if the municipality is in Brazil.
- `Beta 3` is the difference between the expected value of the literacy at any level if the observation is from Paraguay versus not from Paraguay.
- `Paraguay` is a dummy variable which equals "0" if the municipality being observed is in not in Paraguay, and "1" if the municipality is in Paraguay.
- `Beta 4` is the effect that one km distance from rivers would have on the modern litercy level.
- `distance from river` is the distance in km from a river.
- `Beta 5` is the effect that one km distance from the coast would have on the modern literacy level.
- `distance from coast` is the distance in km from the coast.
- `Beta 6` is the effect one cm of annual precipitation might have on the literacy rate.
- `precipitation level` is the the cm of rain there is in an area.
- `Beta 7` is the effect one cm of altitude might have on the literacy rate.
- `altitude` is altitude of an area by cm.
- `error` refers to random error within the OLS regression.

Secondly,

$$\text{annual income} = \alpha_0 + \alpha_1 \cdot \text{distance from Jesuit mission} + \alpha_2 \cdot \text{Paraguay} + \alpha_3 \cdot \text{distance from river} +$$

$$\alpha_4 \cdot \text{distance from coast} + \alpha_5 \cdot \text{precipitation level} + \alpha_6 \cdot \text{altitude} + \text{error}$$

In which:

- `Alpha 0` is the intercept parameter, in which the distance from a Jesuit mission is 0 km.

- `Alpha 1` is the slope parameter, which represents how the percentage of the total population's literacy would change with a one kilometer difference in km.
- `distance from Jesuit mission` is thus the km the municipality is from a Jesuit mission.
- `Alpha 2` is the difference between the expected value of the literacy at any level if the observation is from Paraguay versus not from Paraguay.
- `Paraguay` is a dummy variable which equals "0" if the municipality being observed is in not in Paraguay, and "1" if the municipality is in Paraguay
- `Alpha 3` is the effect that one km distance from rivers would have on the ln of income.
- `distance from river` is the distance in km from a river.
- `Alpha 4` is the effect that one km distance from the coast would have on the ln of income.
- `distance from coast` is the distance in km from the coast.
- `Alpha 5` is the effect one cm of annual precipitation might have on the ln of income.
- `precipitation level` is the the cm of rain there is in an area.
- `Alpha 6` is the effect one cm of altitude might have on the ln of income.
- `altitude` is altitude of an area by cm.
- `error` refers to random error within the OLS regression.

OLS regression is appropriate in this case because the response variable in both equations is numerical and continuous, in contrast to other relations we've explored in the couse where, because the response is a dummy variable, a logistic regression fit better. That being said, there may be an issue with regard to over-fitting or statistically insignificant predictor variables given all the geographic data that is being accounted for, so I'll examine the residual plots in order to ensure there isn't a persistent pattern that points to a systemic rather than random error in the model.

Particular features of the model to note are that the state effects are categorical and thus will have a level effect on the final regression. Distance from the river and coast, as well as precipitation, will affect the overall slope of the linear trend as it is numerical an the effect is measured on a per-unit basis.

4 Results

Within this section, I will convey both graphical and tabulated representations of the models that have been constructed.

Here, you can see two different scatterplots: one that compares the distance from Jesuit missions against the literacy rates (Figure 3: Scatterplot of Distance from Jesuit Mission and Income R$) and another that plots the distance from the missions against the literacy rates (Figure 4: Scatterplot of Distance from Jesuit Mission and Literacy Rate). For both, observations in which the distance exceeded 225 km were taken out because they were outliers that skewed the data.

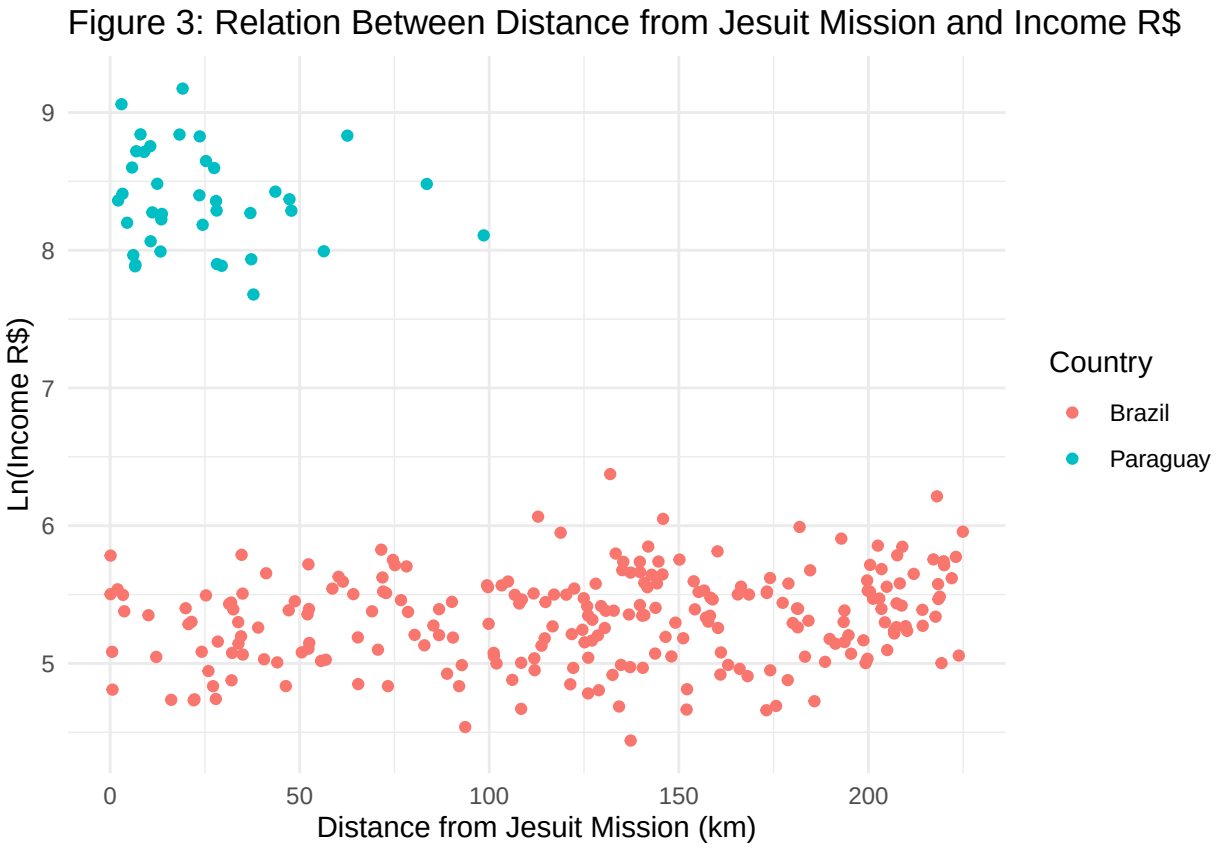
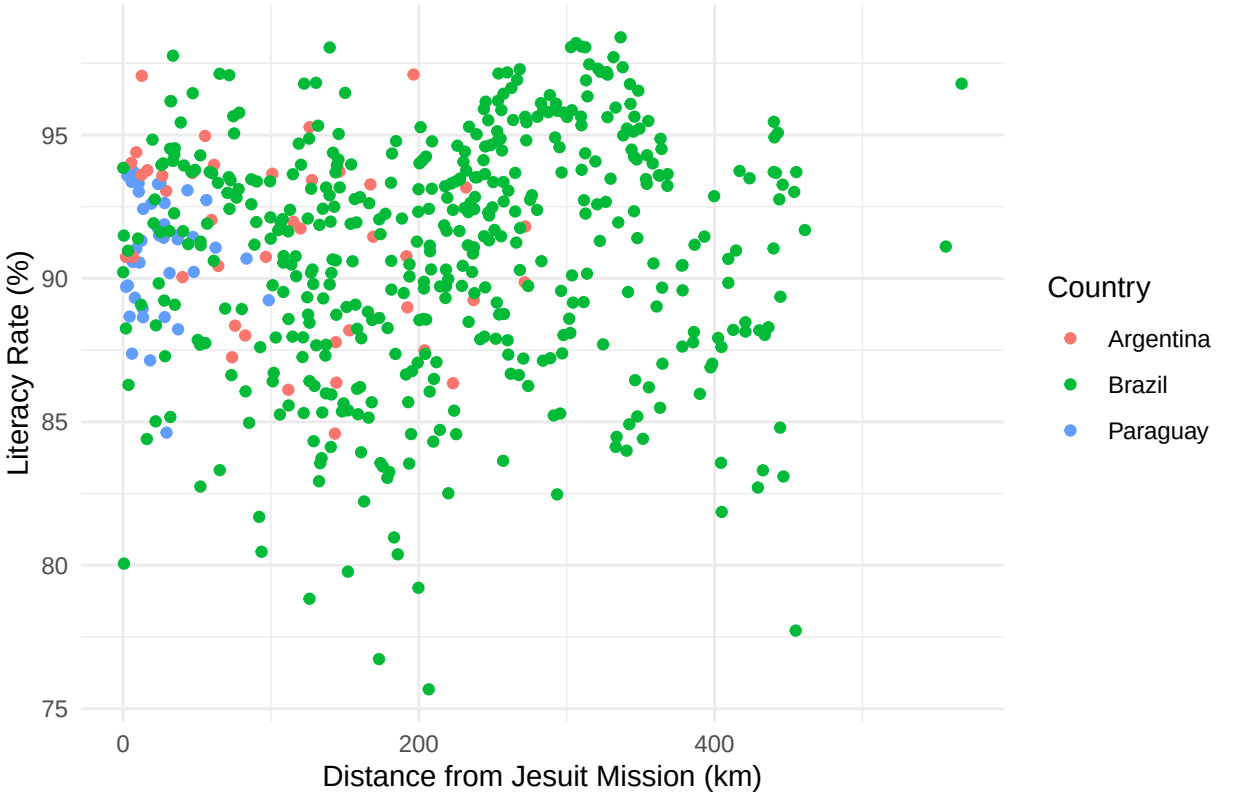


Figure 3: Scatterplot of Distance from Jesuit Mission and Income R$

Figure 4: Relation Between Distance from Jesuit Mission and Literacy Rate



These visual graphs can moreover be translated into multivariate linear regression models. The following tables, Table 2: Regression Analysis for Literacy Rate as Response, and Table 3: Regression Analysis for Ln(Income) as Response have statistics of interest: the beta/alpha values for each of the parameters, the standard error for those values, and the p-value for the individual t-tests done on each variable.

Moreover, Table 4: Comparative Table of Relevant Statistics of Models, compares the p-values of the global F-tests for the overall model as well as the adjusted coefficient of multiple determination, two indicators of a model's appropriateness will be further discussed in **Discussion**.

Table 2: Regression Analysis for Literacy Rate as Response

	Estimate (Beta Values)	Standard Error	P-Value for T-test
Intercept	91.724335	2.712782	< 2e-16
Distance from Mission	-0.004854	0.003417	0.155964
Brazil	-2.877639	0.989400	0.155964
Paraguay	-0.548585	0.915390	0.549230
Coast	-1.062345	0.308446	0.000617
River	-1.801725	0.692352	0.009514
Precipitation	0.004001	0.001545	0.009888
Altitude	-0.002467	-2.307	0.021420

Table 3: Regression Analysis for Ln(Income) as Response

	Estimate (Beta Values)	Standard Error	P-Value for T-test
Intercept	5.862e+00	2.229e-01	< 2e-16
Distance from Mission	-1.249e-03	3.135e-04	0.155964

Paraguay	3.558e+00	8.505e-02	$<$ 2e-16
Coast	-2.327e-01	3.100e-02	2.81e-13
River	-9.429e-02	5.461e-02	0.0849
Precipitation	1.943e-04	1.422e-04	0.1725
Altitude	1.072e-04	8.628e-05	0.2146

Table 4: Comparative Table of Relevant Statistics of Models

	Literacy Rate	Ln(Income)
Adjusted R-squared	0.0432	0.87
P-Value for Global F-Test	6.373e-05	$<$ 2.2e-16

5 Discussion

5.1 Graphical Trends

Interestingly, while Figure 2 appears to portray a relatively similar trend for all the included countries, a gradual trend downward seemingly to indicate a negative relation between distance from a Jesuit mission and the literacy rate at the same rate for the different countries, the data regarding how the ln(income) is affected for the different countries regarding distance from the sites is dramatically different. Paraguay for instance seeems to have overall higher average incomes per municipality than Brazil's, but while Paraguay's (which ae also overall closer to mission sites) seem to have a negative relation with difference from a mission site, Brazil's appears to be either completely null with no relation, or perhaps a very small positive relation—i.e., that being farther from a mission site appears to increase the ln(income) rather than decrease it.

That being said, these two graphs only take into account distance from a mission site as the shole predictor, whereas it is certainly possible other geographical factors such as access to a river or distance from the coast may also influence some of the variation we see with the scatterplot. For instance, perhaps it is more likely that municipalities closer a river might have a higher income on average because of access to resources, a dynamic that might muddy the results a simple comparison of two variables might miss. For this reason, it's important that we do a tabulated OLS analysis that takes into account different geographical data, including: distance from a river, distance from the coast, precipitation and altitude.

5.2 Model Results

I will now analyze the direct results from Table 2 and Table 3, with reference to their p-values for their individual t-tests. First of all, to contextualize why the p-value is important, this value gives an indication in a multivariate context whether any one regression coefficient is able to predict our chosen response variable over the model if this particular predictor wasn't included. For instance, if we're testing the variable `distance from mission`, the p-value for the t-test investigates if that variable actually helps the model predict `literacy rate` or `ln(income)` above the other variables.

The null hypothesis for each coefficient is that it equals 0 (thus, there is no relation between the predictor and the response). A smaller p-value for each individual t-test indicates strong evidence against the null hypothesis that the predictor has no predictive ability on the chosen response variable. The threshold for the p-value is usually about 0.05, such that there isn't enough evidence to indicate the null hypothesis, that there is no relation, false. On the other hand, a global F-test investigates the *entire* model, to see if the total collection of variables are able to create a compelling model which will provide useful information regarding the relations the predictors have with the response variables.

According to Table 2, the distance from a mission appears to be insignificant, in that the p-value is much larger than the significance level. It appears that this variable does not help the model be more accurate more so than a model that doesn't in fact include it, suggesting that the coefficient assigned to it, -0.004854 (indicating for every km increase in distance the literacy rate decreases by 0.004854%, or more helpfully for every 100 km in distance a -0.4854% effect on the literacy rate) may not be accurate. Table 3, measuring the effect the distance from a mission has on the ln of annual income, the effect is measured as -1.249e-03% for every km in distance, so also very miniscle of an effect, but the p-value is much higher than the significance threshold, suggesting there is very little evidence against the hypothesis that this variable aids the model in a substantial way.

However, what's interesting to note is that the overall model, according to the global F-test, appears to in both cases have a p-value below the significance level. Thus, this indicates that though distance from a missionary site may no be a valid explanatory variable for predicting literacy levels or ln of income, there *are* some other variables which may be useful in this regard. It is particularly interesting how distance from the river and coast do not only have relatively more substantial effects on the literacy level according to Table 2 (-1.062345% and -1.801725 respectively), their p-values are also very low, suggesting that the null hypotheses that these relations do not exist have substantial evidence against them. This may suggest that geographical variables are more important to predicting literacy levels than previously realized, and may be a more reliable predictor than distance from missionary sites. In terms of significance, for the literacy level model, distance from the coast, distance from the river, altitude and precipitation are all considered useful predictors. For ln(income), the country that the observation is from along with the coast are significant and considered useful predictors.

5.3 Strengths, Weaknesses and Next Steps

There are crucial strengths and weaknesses to the analysis conducted. In terms of strengths, summarizing the complex relationship that Jesuit missions held towards the communities in terms of distance was a very transformative way to consider the spatial relations. This data makes use of very thorough geographical data that could very well affect both literacy and ln(income)—after all, it makes logical sense that access to a river might be correlated with access to more fertile land, which in turn may increase the average income of a resident. The level of income of a community also relates to the level of literacy of the population, so though this data seems to indicate that there is not enough evidence to reject the null hypothesis that there is no relation between distance from a site and the literacy and income of a community, there is enough to suggest other variables related to the inherent geography of a location may provide the framework for a further analysis, this one focused wholely on access to natural resources and the effect that has on indicators of human capital.

Weaknesses of this data from the outset have been that the observations from different countries come from different years. Though the range from 2000 to 2002 may seem small for the literacy variable, it is still a sizeable difference and creates a degree of uncertainty within the data. This is further dramatic with income, which has a larger year range and also excludes Argentina. It is interesting to note that the further away a municipality was from a Jesuit mission in Brazil, it seemed to increase the literacy rate, suggesting that perhaps the story is not as simple as Jesuit missions increasing human capital.

6 References

- Caicedo, Felipe Valencia, 2018, "Replication Data for: 'The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America' ", https://doi.org/10.7910/DVN/ML1155, Harvard Dataverse, V1
- Caicedo, Felipe Valencia, 2018, "The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America", https://economics.harvard.edu/files/economics/files/ms25323.pdf, Harvard Dataverse, V1

- "Continuous National Household Sample Survey - Continuous PNAD." IBGE, https://www.ibge.gov.br/en/statistics/social/population/16833-monthly-dissemination-pnadc1.ht ml?=&t=o-que-e.
- Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven
- JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com.
- Luckhurst, Toby (2018). John Allen Chau: Do missionaries help or harm? BBC News, https://www.bbc.com/news/world-46336355.
- "OECD Assessment of the Statistical System of Argentina and Key Statistics of Argentina." OECD, http://www.oecd.org/statistics/good-practice-toolkit/countryassessments/OECD-Assessment -of-the-Statistical-System-and-Key-Statistics-of-Argentina.pdf.
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.26.