

## Replication Study: Housing, Health, and Happiness

### Introduction & Summary Statistics








For our final project, we will be conducting a replication study of Matias D. Cattaneo et.al “Housing, Health, and Happiness” 2009. We will begin by introducing the paper’s main research questions and conclusions along with basic data exploration and a summary statistics table of important covariates.

**In a sentence, Matias D. Cattaneo and his peers study the health effects of the *Piso Firme* project, a government-subsidized “slum upgrade” involving the installation of cement floors in households near the Torreon region of Mexico.** In particular, they compare outcome variables of parasitic infestations, diarrhea, anemia, height, weight, and cognitive development against the untreated Durango region across state lines. Moreover, the regression model includes 41 covariates, including both numerical and dummy variables such as parasite count, total income, e.t.c.





































To ensure causality between the Piso Firme program and an increase in household health/happiness, households in the study undergo a pre-intervention check, to ensure that we are using comparable households from both neighborhoods. This is done through comparing variables such as father presence, presence of diseases, respiratory issues, intellectual scores, e.t.c. - there are no significant differences in mean values between control and treatment groups. This implies that improvements can be attributed to the Piso Firme program and the installation of a cement floor.

With respect to data cleaning, there seems to be the presence of missing/null values within all rows of both datasets; researchers seem to impute missing values with ‘zeros’ as many of them are dummy variables and the missing values seem missing at random; ie, imputing with zeroes is a quick and easy solution to a nuanced and complicated data-completeness issue. Finally, you may find below a table of the covariates and outcome variables of the regression model implemented by Cattaneo et al.; other variables were used in pre-intervention studies for causality purposes and not shown below.

### Outcomes:

skim_variable	mean	sd	p25	p75	hist
S_parcount	0.302	0.625	0	0	
S_diarrhea	0.133	0.340	0	0	
S_anemia	0.386	0.487	0	1	
S_mccdts	15.4	20.1	2	20	
S_pbdypct	31.9	25.4	12	47	
S_haz	-0.602	1.11	-1.33	0.100	
S_whz	0.131	1.13	-0.560	0.720	

Covariates:

skim_variable	mean	sd	p25	p75	hist
dpisofirme	4.90e-1	5.00e-1	0	1	
idcluster	2.09e+8	1.37e+8	70001446	350002560	
coord_x	-1.03e+2	5.52e-2	-103.	-103.	
coord_y	2.56e+1	3.71e-2	25.5	25.6	
idmun	2.10e+1	1.37e+1	7	35	
idmza	9.84e+1	2.59e+2	12	38	
S_age	1.34e+1	1.43e+1	2	27	
S_childma	9.65e-1	1.83e-1	1	1	
S_childmaage	2.94e+1	9.43e+0	23	33	
S_childmaeduc	6.69e+0	2.84e+0	6	9	
S_childpa	7.80e-1	4.14e-1	1	1	
S_childpaage	3.23e+1	1.02e+1	26	36	
S_childpaeduc	6.74e+0	3.20e+0	6	9	
S_HHpeople	5.63e+0	2.17e+0	4	6	
S_rooms	2.04e+0	1.09e+0	1	3	
dtrriage_5_02_male	1.05e-2	1.02e-1	0	0	
dtrriage_5_35_male	1.55e-2	1.24e-1	0	0	
dtrriage_5_68_male	1.67e-2	1.28e-1	0	0	
dtrriage_5_911_male	1.66e-2	1.28e-1	0	0	
dtrriage_5_02_female	9.26e-3	9.58e-2	0	0	
dtrriage_5_35_female	1.64e-2	1.27e-1	0	0	
dtrriage_5_68_female	1.51e-2	1.22e-1	0	0	
dtrriage_5_911_female	1.51e-2	1.22e-1	0	0	
S_waterland	9.73e-1	1.63e-1	1	1	
S_waterhouse	5.23e-1	4.99e-1	0	1	
S_electricity	9.90e-1	1.02e-1	1	1	
S_milkprogram	6.89e-2	2.53e-1	0	0	
S_foodprogram	2.88e-2	1.67e-1	0	0	
S_seguropopular	1.42e-2	1.18e-1	0	0	
S_hasanimals	4.93e-1	5.00e-1	0	1	
S_animalsinside	1.93e-1	3.94e-1	0	0	
S_garbage	8.19e-1	3.85e-1	1	1	
S_washhands	3.74e+0	1.49e+0	3	4	
S_incomepc	1.06e+3	3.47e+3	533.	1086.	
S_cashtransfers	1.31e+1	3.42e+1	0	0	
S_assetspc	2.14e+4	7.28e+3	16844.	26313.	

### Model Assumptions

The authors implemented a least squares regression combined with a sandwich method for clustering standard errors among each household. The mathematical formulation of the model is as follows:

$$Y \sim X\hat{\beta}$$

$$\hat{\beta} = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$$

where  $\Omega$ , the clustering element is defined as:

$$\Omega = \frac{n-1}{n-p} \frac{c}{c-1} \sum_{j=1}^c (X_j^T \hat{e}_j \hat{e}_j^T X_j)$$

Here,  $n$  and  $p$  represent the number of observations and features respectively, while  $c$  denotes the number of clusters.  $X$  is the independent variable dataset, and  $e$  is the observed residual upon running an OLS regression.

The model requires four primary assumptions:

1. Linearity
2. Identical error distribution within clusters
3. No autocorrelation

These statements can be summed up mathematically:

1.  $Y \sim X\hat{\beta}$
2.  $E(e_j) \sim N(0, \sigma^2) \text{ i. i. d.}$

### Main Result Replication

Three primary forms of this model were constructed, which vary in their choice of dependent variables, regressing all models against each dependent variable of note. The first model contains no independent variables beyond whether a household/individual was part of the Piso Firme program. The second model includes demographic and health data of the target. The third model includes the data of the previous models, as well as additional data on a household/individual's involvement in relevant state social programs.

For each dependent variable, each model is built and the coefficient, standard error (with \* symbols denoting statistical significance), and standard error normalized to the mean, are recorded.

The first hypothesis tested by the authors seeks to confirm that the Piso Firme program was successful in installing concrete floors in the target region at a higher rate than occurred in regions where the program was not active. All dependent variables returned high statistical significance, confirming this fact.

The second series of dependent variables tested by the authors involved an attempt to quantify the effect on the health and cognitive development of children that the implementation of the program had. While not all variables returned statistical significance, particularly some of those involving cognitive testing scores, they did find strong evidence that children involved in the Piso Firme program were generally at lower risk of certain health issues.

The third table generated evaluated the program's effect on maternal happiness rates. These models were extremely effective and demonstrated a strong association between involvement in the Piso Firme program and a higher maternal quality of life.

Below are our replications of the models. These are nearly identical to the tables reported in the study. Some standard error values differ from those given by up to .005, although most are as near as .001. This is because, even after testing multiple R packages, we found that no sandwich estimating R package for clustered errors perfectly replicates the algorithm implemented in the Stata function used by the authors. The values shown below come from the package that we found maximized performance.

### 1. *Estimated Program Impact on Cement Floors*

<b>Replication of Table 4 - Regressions of Cement Floor Coverage Measures on Program Dummy</b>				
<b>Dependent Variable</b>	<b>Control Group Mean (Standard Deviation)</b>	<b>Model 1</b>	<b>Model 2</b>	<b>Model 3</b>
<b>Share of rooms with cement floors</b>	0.717 (0.401)	0.211 [0.021]*** 29.533	0.214 [0.019]*** 29.870	0.216 [0.019]*** 30.185
<b>Cement floor in kitchen</b>	0.657 (0.540)	0.266 [0.024]*** 40.467	0.267 [0.022]*** 40.652	0.271 [0.022]*** 41.361
<b>Cement floor in dining room</b>	0.702 (0.528)	0.219 [0.027]*** 30.882	0.219 [0.025]*** 31.453	0.219 [0.025]*** 32.128
<b>Cement floor in bathroom</b>	0.791 (0.495)	0.219 [0.023]*** 14.623	0.219 [0.019]*** 15.144	0.219 [0.019]*** 15.557
<b>Cement floor in bedroom</b>	0.653 (0.558)	0.219 [0.221]*** 37.858	0.219 [0.221]*** 38.696	0.219 [0.221]*** 38.610

*Notes: Regressions computed using survey information. Model 1: no controls; Model 2: age, demographic, and health-related controls; Model 3: age, demographic, health-habits and public social programs controls. Reported results: estimated coefficient, clustered standard error at census-block level in brackets (136 clusters), and 100 x coefficient/control mean*

*\*\*\* Significantly different from 0 at 1 percent level*

*\*\* Significantly different from 0 at 5 percent level*

*\* Significantly different from 0 at 10 percent level*

*Source: 2005 Survey*

## 2. Estimated Program Impact on Child Health

Replication of Table 5 – Regressions of Children's Health Measures on Program Dummy				
Dependent Variable	Control Group Mean (Standard Deviation)	Model 1	Model 2	Model 3
Parasite Count	0.333 (0.876)	-0.065 [0.032]** -19.650	-0.065 [0.031]** -19.401	-0.065 [0.032]** -19.252
Diarrhea	0.142 (0.471)	-0.018 [0.009]* -12.909	-0.020 [0.009]** -13.985	-0.018 [0.009]* -12.945
Anemia	0.426 (0.670)	-0.085 [0.028]*** -19.884	-0.080 [0.027]*** -18.802	-0.082 [0.027]*** -19.284
MacArthur Communicative Development Test Score	13.354 (27.780)	4.031 [1.540]** 30.182	5.652 [1.642]*** 42.325	5.557 [1.641]*** 41.609
Picture Peabody Vocabulary Test percentile score	30.656 (35.337)	2.668 [1.688] 8.702	3.206 [1.430]** 10.460	3.083 [1.410]** 10.058
Height-for-age z-score	-0.605 (1.527)	0.007 [0.043] -1.230	-0.001 [0.038] 0.242	0.002 [0.039] -0.361
Weight-for-height z-score	0.125 (1.558)	0.001 [0.034] 1.163	-0.005 [0.036] -4.569	-0.011 [0.037] -9.188

Notes: Regressions computed using survey information. Model 1: no controls; Model 2: age, demographic, and health-related controls; Model 3: age, demographic, health-habits and public social programs controls. Reported results: estimated coefficient, clustered standard error at census-block level in brackets (136 clusters), and 100 x coefficient/control mean

\*\*\* Significantly different from 0 at 1 percent level

\*\* Significantly different from 0 at 5 percent level

\* Significantly different from 0 at 10 percent level

Source: 2005 Survey

### 3. Estimated Program Impact on Maternal Happiness

Replication of Table 6—Regressions of Satisfaction and Maternal Mental Health Measures on Program Dummy				
Dependent Variable	Control Group Mean (Standard Deviation)	Model 1	Model 2	Model 3
Satisfaction with floor quality	0.508 (0.661)	0.221 [0.023]*** 43.575	0.227 [0.024]*** 44.624	0.226 [0.025]*** 44.483
Satisfaction with house quality	0.604 (0.664)	0.086 [0.021]*** 14.289	0.082 [0.021]*** 13.568	0.078 [0.022]*** 12.979
Satisfaction with quality of life	0.593 (0.661)	0.114 [0.023]*** 19.258	0.112 [0.022]*** 18.861	0.111 [0.023]*** 18.731
Depression scale (CES-D scale)	18.488 (12.365)	-2.484 [0.632]*** -13.434	-2.553 [0.582]*** -13.809	-2.504 [0.573]*** -13.543
Perceived stress scale (PSS)	16.457 (9.360)	-1.771 [0.441]*** -10.762	-1.774 [0.403]*** -10.780	-1.749 [0.413]*** -10.630

Notes: Regressions computed using survey information. Model 1: no controls; Model 2: age, demographic, and health-related controls; Model 3: age, demographic, health-habits and public social programs controls. Reported results: estimated coefficient, clustered standard error at census-block level in brackets (136 clusters), and 100 x coefficient/control mean

\*\*\* Significantly different from 0 at 1 percent level

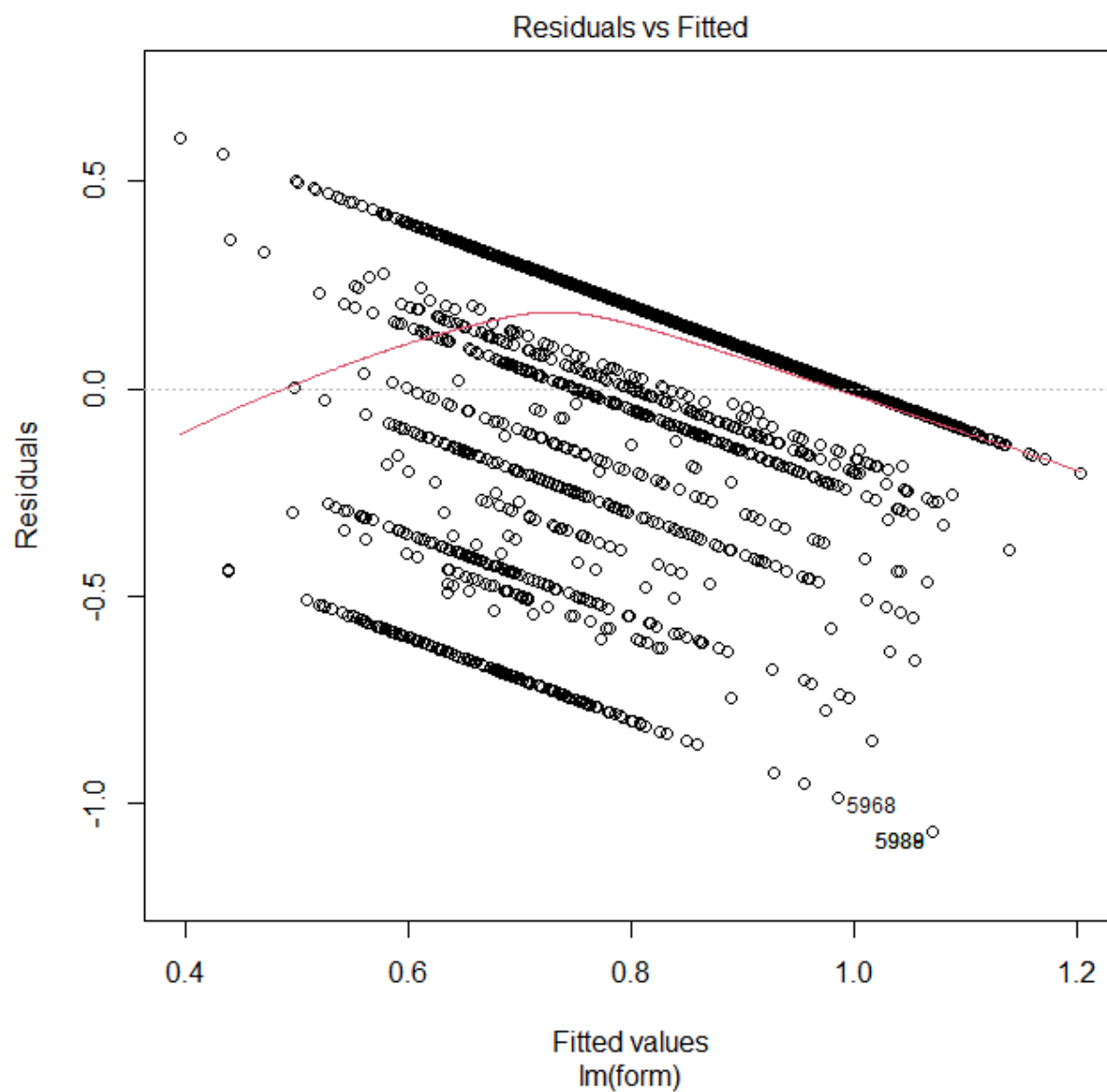
\*\* Significantly different from 0 at 5 percent level

\* Significantly different from 0 at 10 percent level

Source: 2005 Survey

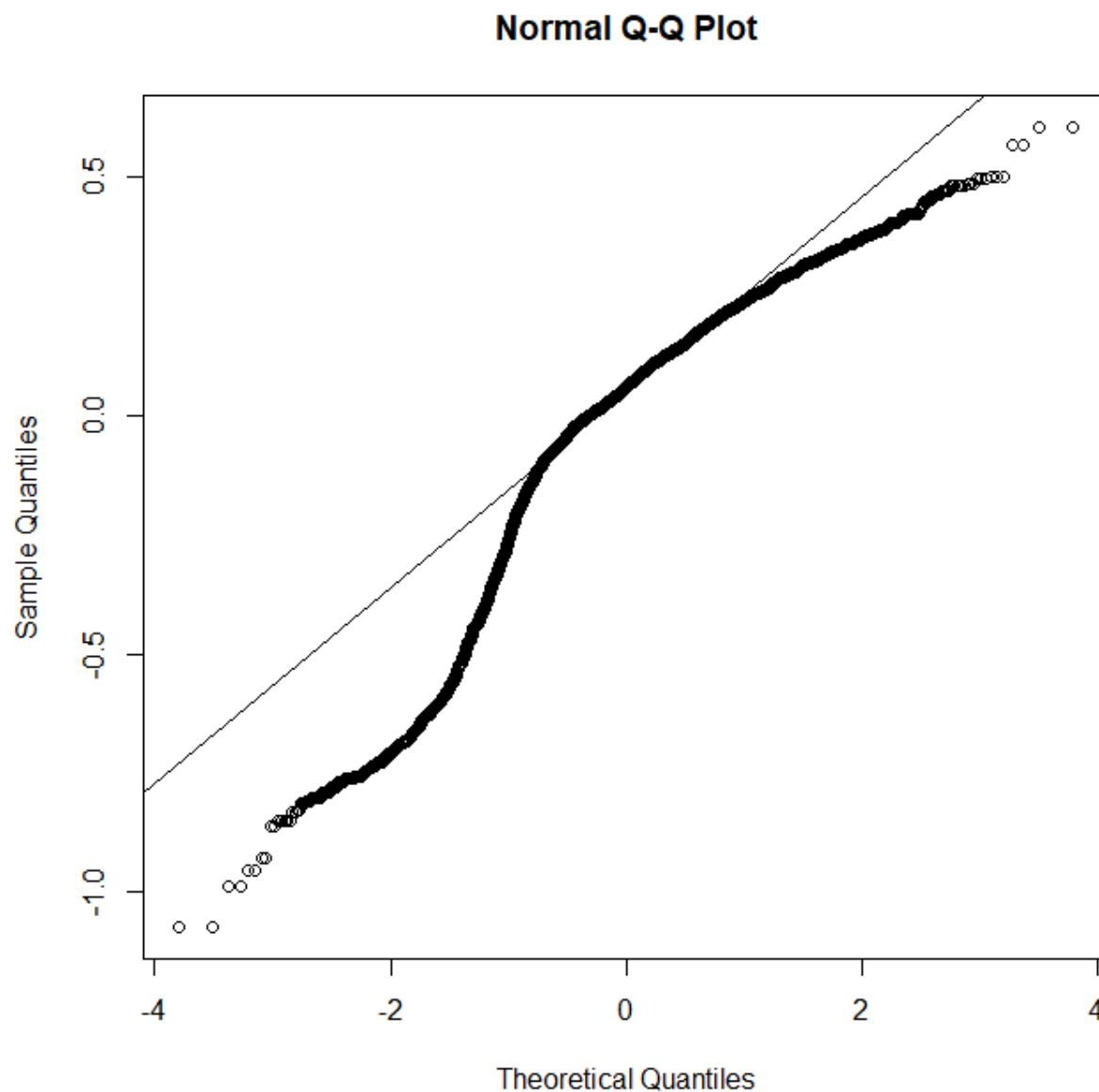
#### Criticism of Model Assumptions

In the interest of brevity, model assumptions will be checked using the most complex model (model 3) with the dependent variable being ‘Share of Rooms with a Cement Floor’.



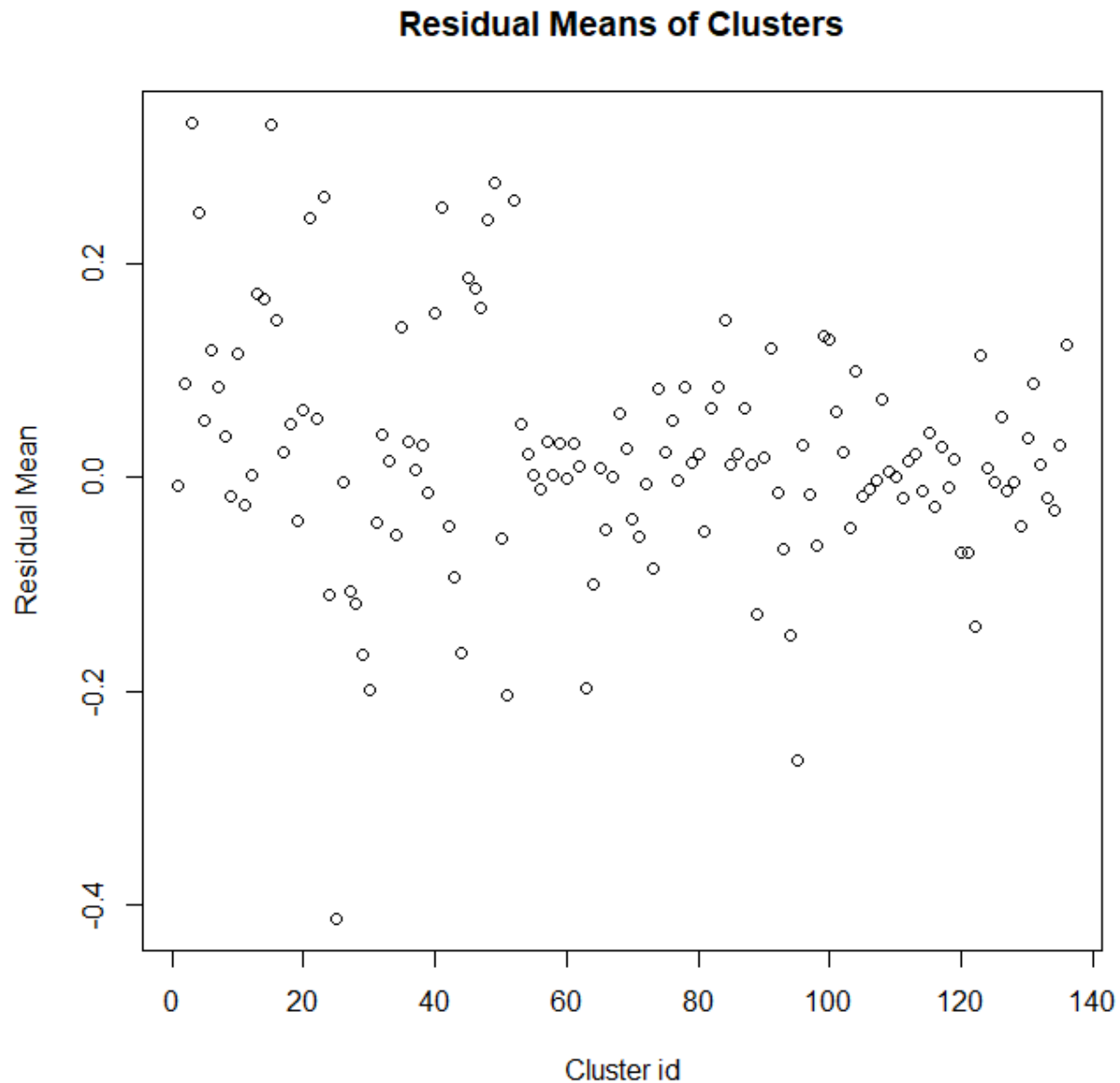
This is far from a linear relationship. There are visible trends which the model may pick up on, but this model is certainly not a good fit.

The second assumption, involving the distribution of residuals within each cluster, we first examine the qq-plot:



While not technically a requirement for this model, this poor alignment indicates that residuals may not be normally distributed, even within clusters. Clusters are small, with some containing just 1 or 2 observations, making it difficult to perform a robust test of normality. Instead, we calculate the mean of residuals within each cluster, to at least check for residual expectation of zero:





As expected, given the small sizes of some clusters, we see some outlying data points. That said, the residuals as a whole appear effectively centered at zero. This assumption check remains incomplete, and given the lack of linearity seen before, it would not be surprising if this check would fail, were more data available to examine cluster error distributions.

On top of the aspects discussed above, the most glaring issue with the methods employed by the authors is seen in the choice to impute the (many) NA values with zeroes, accompanied with a dummy variable to denote incomplete observations. Many of the values being replaced by zero fall in features with non-trivial distributions, making this approach suspect. There is concern that this may have negatively affected the fit of the model. This will be further examined in the Re-analysis section of this paper.

## Robustness Checks

Next, it is to be shown that our metrics pass the same robustness checks as the author's originally published experiments. First, we address causality concerns exploring the possibility of improved living conditions and overall family happiness as the result of other social programs - during the period of *Piso Firme*, the Mexican government . Second, we examine the pathways by which *Piso Firme* directly improves the quality of life for residents in the State of Coahuila. Finally, we validate the utilization of clustered standard errors for our experimental linear regressions.

### *A. Potential Bias from Other State and Local Programs*

To definitively conclude that improvements seen in the State of Coahuila are a result of the *Piso Firme* public program and not a result of other programs implemented in Coahuila but not Durango, we re-apply the linear regressions and test to see if there are significant differences between our treatment (1) and control (0) groups with respect to dependent variables concerning health, asset value, and income.

Before we discuss the results of this first robustness check, note that we are still constrained under the assumptions mentioned in the beginning of this replication study. Let's first re-note these assumptions below:

- Samples are balanced across a large number of socioeconomic, health, and demographic characteristics prior to *Piso Firme*
- Trends in health and socioeconomic indicators prior to *Piso Firme* were the same for both control and treatment groups
- No programmatic or policy differences between treatment & control aside from the cement floor installation granted by *Piso Firme*
- Explicitly controlled for the effects of other social programs such as the cash transfer program, nutrition assistance programs, etc.

From the results of our table below, the treatment dummy variable is not significantly associated with any of the illness measurements. Since there is no detectable difference in 'Respiratory diseases', 'Skin diseases', or 'Other diseases'; i.e., diseases not affected by the presence of a cement or dirt floor, this implies that our previous experiment accurately captures the impact effects of the installation of cement floors in Coahuila. In other words, diseases that we don't expect to be affected show no statistical difference between treatment and control groups, validating the premises of our model validation.

Replication of Table 7 - Robustness Checks				
Dependent Variable	Control Group Mean (Standard Deviation)	Model 1	Model 2	Model 3

<b>Respiratory diseases</b>	0.355 (0.667)	0.021 [0.019] 6.000	0.019 [0.018] 5.420	0.017 [0.019] 4.892
<b>Skin diseases</b>	0.101 (0.418)	0.001 [0.012] 1.027	0.003 [0.012] 2.635	0.002 [0.012] 2.354
<b>Other diseases</b>	0.041 (0.283)	0.006 [0.009] 14.076	0.007 [0.009] 16.459	0.007 [0.009] 15.983
<b>Installation of Cement floor</b>	0.530 (0.576)	0.373 [0.028]*** 70.406	0.370 [0.029]*** 69.814	0.372 [0.028]*** 70.300
<b>Construction of sanitation facilities</b>	0.105 (0.409)	-0.021 [0.017] -19.489	-0.021 [0.016] -19.647	-0.020 [0.016] -18.538
<b>Restoration of sanitation facilities</b>	0.046 (0.289)	-0.003 [0.014] -5.535	-0.003 [0.014] -6.463	-0.003 [0.013] 0.575
<b>Construction of ceiling</b>	0.157 (0.525)	0.029 [0.024] 18.667	0.023 [0.024] 14.411	0.019 [0.023] 12.287
<b>Restoration of walls</b>	0.108 (0.444)	0.014 [0.017] 13.198	0.219 [0.221] 0.0232	0.219 [0.221] 0.0232
<b>Any house expansion (excluding installation of cement floors)</b>	0.276 (0.636)	0.041 [0.033] 14.729	0.032 [0.032] 11.501	0.034 [0.031] 12.178
<b>Log of self-reported rental value of house</b>	5.906 (1.070)	0.037 [0.041] 0.635	0.048 [0.032] 0.818	0.051 [0.032] 0.872
<b>Log of self-reported sale value of house</b>	10.469 (1.639)	-0.020 [0.098] -0.193	-0.006 [0.079] -0.054	-0.006 [0.077] -0.054
<b>Log total income of mothers of children 0-5 years</b>	7.791 (0.821)	-0.037 [0.064] -0.480	-0.034 [0.065] -0.436	-0.029 [0.066] -0.374
<b>Log total income of</b>	8.121	-0.016	-0.005	0.219

<b>fathers of children 0-5 years</b>	(0.853)	[0.028] -0.194	[0.027] -0.064	[0.026] 0.016
<b>Total consumption per capita</b>	726.525 (1702.537)	22.890 [53.682] 3.151	30.770 [58.954] 4.235	34.368 [58.466] 4.731

*Notes: Regressions computed using survey information. Model 1: no controls; Model 2: age, demographic, and health-related controls; Model 3: age, demographic, health-habits and public social programs controls. Reported results: estimated coefficient, clustered standard error at census-block level in brackets (136 clusters), and 100 x coefficient/control mean*

*\*\*\* Significantly different from 0 at 1 percent level*

*\*\* Significantly different from 0 at 5 percent level*

*\* Significantly different from 0 at 10 percent level*

*Source: 2005 Survey*

### **B. Pathways**

With an estimated market value of \$150, or half of the average monthly income, researchers want to be certain that these measured increases in quality of life are due to the installation of cement floors and not the financial alleviation that the free installation of a cement floor may bring.

Within our replication of Table 7, notice that covariates related to household income and wealth and household value such as rental/sale values of properties and the log-normalized incomes of both mothers & fathers have no significant correlation with any of our control models. Surprisingly, families receiving *Piso Firme* do not seem to be motivated in other household renovations. Moreover, there exists no statistically significant difference in incomes or house value between families receiving *Piso Firme* and those families who haven't; i.e., the income and house value between treatment and control groups show no significant difference.

### **C. Specification test for Clustered Standard Errors**

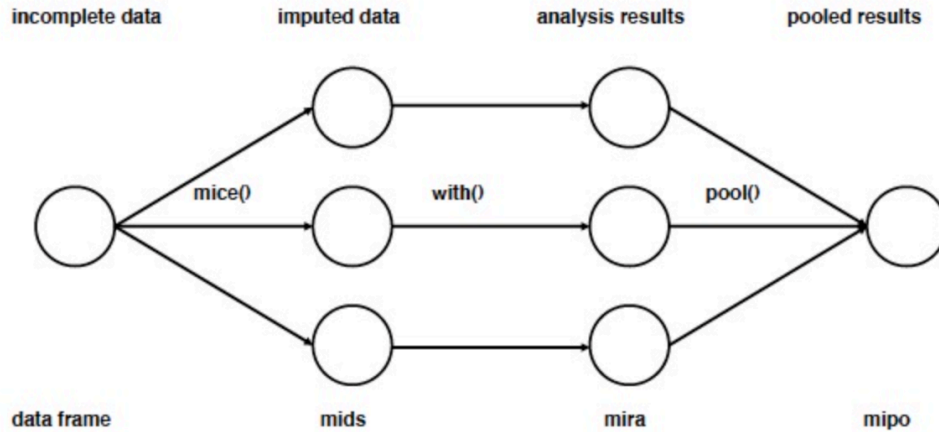
Intuitively, our clustering structure is a natural extension to our sampling methodology; in a sentence, we believe that families sampled within the same id-cluster have correlated measurements. However, proponents of this clustering scheme may argue that correlation amongst clustered errors would invalidate our assumptions of cluster independence. Through the implementation of a between-cluster Moran Hypothesis Test, **we fail to reject the null hypothesis of no-spatial serial correlation for each regression estimated in this paper, and thus conclude that there is insufficient evidence to indicate correlation between our 136 clusters.**

### **Re-analyzation**

Finally, we are proposing a different methodology of missing data imputation for our re-analyzation step; the author's imputation of all missing values with '0' potentially results in errors and a misrepresentation of our sampled data and **incorrect conclusions**. Our proposed missing data imputation is through MICE, or multiple chained equations. Note that implementing

MICE when data are not MAR could result in biased estimates. In the remainder of this paper, we assume that the MICE procedures are used with data that are MAR.

*Fig 1. MICE Imputation Methodology*



Many of the initially developed multiple imputation procedures assumed a large joint model for all of the variables, such as a joint normal distribution. In large datasets, with hundreds of variables of varying types, this is rarely appropriate. MICE is an alternative, flexible approach to these joint models. In fact, MICE approaches have been used in datasets with thousands of observations and hundreds (e.g. 400) of variables; in the MICE procedure a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution, with, for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression. The steps for our imputation is detailed below:

**Step 1:** Replace (or impute) the missing values in each variable with temporary "placeholder" values derived solely from the non-missing values available for that variable. For example, replace the missing age value with the mean age value observed in the data, replace the missing income values with the mean income value observed in the data, etc.

**Step 2:** Set back to missing the "place holder" imputations for the age variable only. This way, the current data copy contains missing values for age, but not for income and gender.

**Step 3:** Regress age on income and gender via a linear regression model (though it is possible to also regress age on only one of these variables); to be able to fit the model to the current data copy, drop all the records where age is missing during the model fitting process. In this model, age is the dependent variable and income and gender are the independent variables.

**Step 4:** Use the fitted regression model in the previous step to predict the missing age values. (When age will be subsequently used as an independent variable in the regression models for other variables, both the observed values of age and these predicted values will be used.) The article doesn't make it clear that a random component should be added to these predictions.

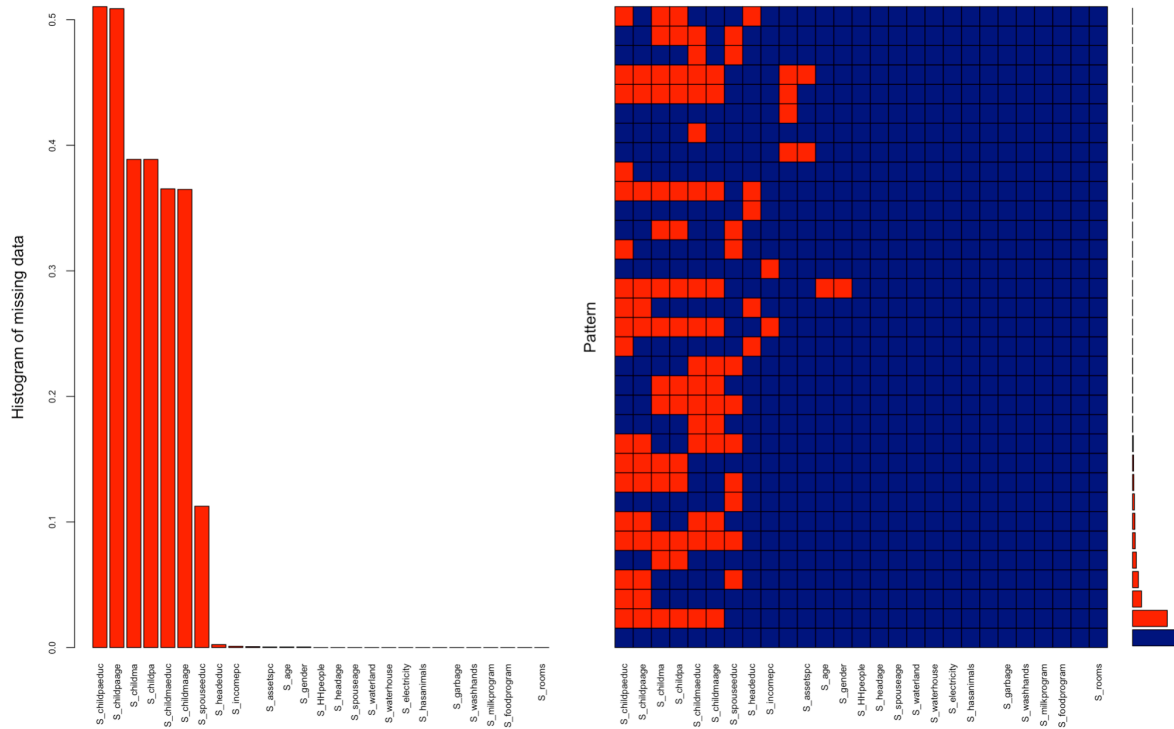
**Step 5:** Repeat Steps 2–4 separately for each variable that has missing data, namely income and gender. [3]

*Fig 2 . Imputed Features by Descending Order of ‘% Missing’*

<b>Imputed Feature</b>	<b>Description</b>
S_childpaeduc	Father’s years of schooling - if present
S_childpaage	Father’s age - if present
S_childma	Presence of mother in household (binary)
S_childpa	Presence of father in household (binary)
S_childmaeduc	Mother’s years of schooling - if present
S_spouseeduc	Spouse’s years of schooling
S_headeduc	Head of household’s years of schooling
S_incompec	Total household income per capita
S_assetpc	Total value of household assets per capita
S_age	Age of survey respondent

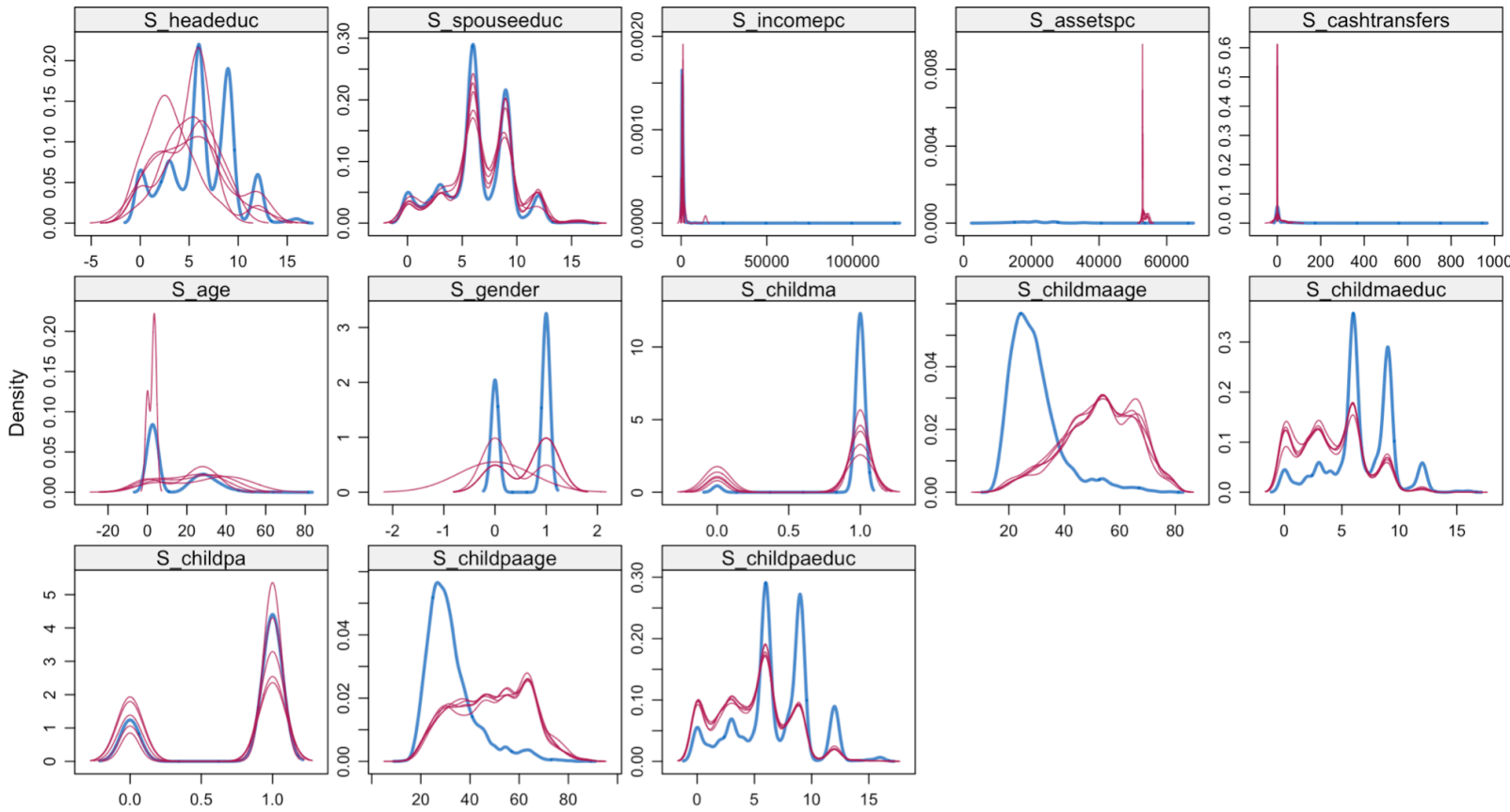
*Source: 2005 Survey*

*Fig 3. Missingness of Replication Study Dataset*



From Figure 2, we can see that **the majority of missing values is present in the columns describing family variables** such as years of schooling for the spouse, husband, . These covariates are present in our control regression models. After employing the MICE imputation technique , we see that the large percentages of missing data have been imputed for certain covariates through chained regressions; Figure 3 provides an accurate depiction of our missingness and we can see the imputed values below.

*Fig 4. Distribution of MICE Imputed Covariates*



There are a couple of issues when Imputing through MICE: we immediately notice **class imbalance issues** in the features of ‘S\_incomepc’, ‘S\_asset\_pc’ and ‘S\_childmaage’, where it seems that our chained regression imputations have a different distribution than the seen data. **We suspect this imputation to be a result of low sample sizes (~1000), but it should not have a major negative bias on our model fittings due to the low feature importance of these covariates.**

After imputing and rerunning the same experiments, we can compare both models through the metrics of AIC and R-squared values. One small nuance here was the selection of comparison metrics between the replicated model and our MICE-imputed model. Due to the number of independent features remaining consistent across all six experiments, we should compare them through  $R^2$  and AIC to determine a relative fitting of our linear regressions, as opposed to metrics such as adjusted  $R^2$  or BICC.

### ***I. Comparison of ‘Regressions of Cement Floor Coverage Measures on Program Dummy***

Metric ( $R^2$ )	Model 1	Model 2	Model 3	MICE Imputed	MICE Imputed	MICE Imputed
---------------------	---------	---------	---------	-----------------	-----------------	-----------------



(AIC)				Model 1	Model 2	Model 2
Share of rooms with cement floors	0.119 2283.755	0.197 1692.723	0.197 1684.726	0.120 2283.755	0.196 1698.045	0.1976 1690.855
Cement floor in kitchen	0.105 6251.794	0.162 5838.490	0.164 5822.142	0.105 6251.794	0.160 5849.871	0.1630 5833.32
Cement floor in dining room	0.08 5957.345	0.126 5604.120	0.130 5585.337	0.075 5957.345	0.126 5603.896	0.129 5584.931

Notes: Regressions computed using survey information. Model 1: no controls; Model 2: age, demographic, and health-related controls; Model3: age, demographic, health-habits and public social programs controls.

## II. Comparison of ‘Regressions of Children’s Health Measures on Program Dummy’

Metric ( $R^2$ ) (AIC)	Model 1	Model 2	Model 3	MICE Imputed Model 1	MICE Imputed Model 2	MICE Imputed Model 2
Parasite Count	0.002 5861.146	0.038 5812.478	0.038 5819.156	0.002 5861.146	0.038 5809.68	0.038 5814.842
Diarrhea	0.0005 2746.022	0.032 2679.122	0.034 2678.462	0.0005 2746.022	0.033 2675.823	0.033 2678.701
Anemia	0.007 5179.236	0.056 5055.323	0.056 5059.895	0.007 5179.236	0.057 5047.517	0.057 5049.693

Notes: Regressions computed using survey information. Model 1: no controls; Model 2: age, demographic, and health-related controls; Model 3: age, demographic, health-habits and public social programs controls.

From the table above, we can see that the  $R^2$  and AIC are similar across both ‘0’-imputation and MICE imputation through chained linear regressions; as stated above, we suspect this is due to the low contribution that these imputed covariates provide in explaining the dependent variable. **It would be fair to say that imputing through Multiple Chained Equations (MICE) produces no noticeable effect on the fitment of our regression models.**

### References

1. Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, 20(1), 40–49. <https://doi.org/10.1002/mpr.329>
2. Kassambara. (2018, November 3). *Regression Model Accuracy Metrics: R-square, AIC, BIC, Cp and more*. STHDA. Retrieved May 7, 2022, from <http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/>
3. Matias D. Cattaneo, Sebastian Galiani, Paul J. Gertler, Sebastian Martinez and Rocio Titiunik, “Housing, Health, and Happiness”, *American Economic Journal: Economic Policy*, 2009