Stat 230A: Replication of Housing, Health, and Happiness

Group 10: Yiyun Gong and Daniel Huang

May 13, 2022

1 Introduction

1.1 Paper Summary

To explore the effect that housing plays health and welfare, Matias Cattaneo et al. [1] evaluated the impact of floor quality, a particular aspect of housing, on the health of young children and the happiness of adults by investigating the effect of a large-scale Mexican government program Piso Firme (launched in 2000) with control-treatment groups. Piso Firme offered concrete cement flooring to low-income households with dirt floors to replace with. By the time this study was conducted (2005), cement floors were installed in an estimated 3 million houses that reported dirt floors in 2000.

1.2 Experiment Outline

The study was conducted by examining the two cities with similar socioeconomic demographics (justified through examining socioeconomic factors before and after the program). One city, Torreón, was a participant in the Piso Firme program while the other two cities, Gómez Palacio and Lerde, were not. One major goal of the study was to examine the influence of flooring on the health of young children and the happiness of their mothers. The health of young children was measured through decreases in the incidence of parasitic infestations, diarrhea, prevalence of anemia, and improvement in children's cognitive development. The happiness of their mothers was measured by satisfaction with their housing and quality of life (welfare), depression score and perceived stress scales. These measurements were further divided into smaller pieces of measurement in the study. The data was collected through surveys conducted in 2005 and census information from 2000 from each household, including household demographic information, socioeconomic status, housing infrastructure, health outcomes, cognitive development of children etc.

1.3 Data Description

The data was collected from surveying 3000 cross-sectional households, equally split across treatment and control groups in the spring of 2005. This data is then supplemented with the census data from 2000 in order to control for confounding factors other than the Piso Firme program. The household dataset contains 2783 household records with 78 variables, and the individual dataset contains 6693 individual records with 89 variables. The household dataset contains household level information, such as the presence of cement floors in certain rooms of the house. The individual data contains individual level information, such as depression and perceived stress scores. Both tables are used in the analysis of the paper.

The impact of the cement floor was analyzed by regressing variables of interest (children's health, adult's mental health) with dummy variables to indicate control/treatment group and a large set of control variables. The researchers found that children in households that were covered by Piso Firme have fewer parasites, lower incidence of diarrhea and anemia, and better cognitive development. Adults in households that were covered by Piso Firme were found to have significantly lower depression and perceived stress scales. In conclusion, the study conducted by Matias Cattaneo et al., found that replacing dirt floors with cement floors can significantly improve the health of young children and the happiness level in adults.

1.4 Tables

Here we replicate some tables from the paper, as well as include some of our own in order to get a better understanding of the data. These tables include sample sizes of variables of interest from the household and individual datasets, summary statistics to examine the distributions of the treatment and control groups respectively across the datasets, as well as independent variables sample sizes and difference of means.

1.4.1 Sample Sizes

Below are the tables containing the sample sizes for the treatment and control groups respectively for each dependent variable. The table in the paper combines variables from the household and individual datasets, while we leave them separate here:

Household Outcome Data Sample Sizes

Variable	**Control**	**Treatment**
Share of Rooms with Cement Floors	1393	1362
Cement Floor in Kitchen	1393	1362
Cement Floor in Dining	1393	1362
Cement Floor in Bathroom	1393	1362
Cement Floor in Bedroom	1393	1362
Satisfaction with Floor Quality	1393	1362
Satisfaction with House Quality	1393	1362
Satisfaction with Life Quality	1393	1362
Depression Scale (CES-D Scale)	1388	1354
Perceived Stress Scale (PSS)	1387	1359
Installation of Cement Floor	1392	1362
Construction of Sanitation Facilities	1390	1362
Restoration of Sanitation Facilities	1391	1362
Construction of Ceiling	1392	1361
Restoration of Walls	1392	1362
Any Improvement	1393	1362
Log of Self-Reported Rental Value of House	1285	1284
Log of Self-Reported Sale Value of House	1223	1239
Total Consumption per Capita	1391	1360

Individual Outcome Data Sample Sizes

Variable	Control	Treatment
Parasite Count	1566	1528
Diarrhea	2105	1930
Anemia	1951	1768
MacArthur Communicative Development Test Score	302	291
Picture Peabody Vocabulary Test Percentile Score	817	757
Height-for-age Z-score	2053	1865
Weight-for-height Z-score	2058	1881
Respiratory Diseases	2107	1930
Skin Diseases	2106	1926
Other Diseases	2106	1930
Log Total Income of Mothers of Children 0-5 Years	301	247
Log Total Income of Fathers of Children 0-5 Years	1000	1026

Additionally, the authors included a table of sample sizes for the independent variables as well as the difference in means between the control and treatment groups. We replicate that here for reference.

Independent Variables Sample Sizes and Difference of Means

Variable	Obs_trt	Mean_trt	Obs_ctr	Mean_ctr	Mean_diff
Household Demographics					
Num of Household Members	1362	5.320	1393	5.374	-0.054
Head of Household Age	1362	37.537	1393	37.120	0.417
Spouse Age	1362	29.645	1393	28.772	0.873
Head of Household Years of Educ.	1360	6.128	1391	6.408	-0.28
Spouse Years of Educ.	1207	6.338	1211	6.479	-0.141
Characteristics of children aged 0-5					
Age	1940	2.643	2112	2.579	0.064
Male (=1)	1940	0.492	2112	0.517	-0.025
Mother Present (=1)	1940	0.968	2112	0.964	0.004
Mother Age	1861	27.383	1992	27.465	-0.082
Mother Years of Educ.	1859	7.059	1992	6.910	0.149
Father Present (=1)	1940	0.797	2112	0.763	0.034
Father Age	1480	30.368	1525	30.632	-0.264
Father Years of Educ.	1476	6.839	1519	7.153	-0.314
Housing characteristics					
Num of Rooms	1362	2.080	1393	1.981	0.099
Water Connection (=1)	1362	0.970	1393	0.977	-0.007
Water Connection in House (=1)	1362	0.511	1393	0.546	-0.035
Electricity (=1)	1362	0.985	1393	0.993	-0.008
Share of Cement Floors 2000	1362	0.330	1393	0.327	0.003
Hygienic environment					
Has Animals on Land (=1)	1362	0.517	1393	0.480	0.037
Has Animals Inside (=1)	1362	0.192	1393	0.190	0.002
Garbage Collection (=1)	1362	0.799	1393	0.845	-0.046
Num of Times Washed Hands	1362	3.754	1393	3.716	0.038
Economic characteristics					
Income Per Capita	1361	1024.703	1391	1051.676	-26.973
Value of Assets Per Capita	1361	22393.733	1393	22032.32	361.413
Public social programs					
Cash Transfers Per Capita from Gov.	1361	16.187	1392	12.604	3.583
Beneficiary of Milk Supplement Program (=1)	1362	0.060	1393	0.082	-0.022
Beneficiary of Gov. Food Program (=1)	1362	0.037	1393	0.022	0.015

Notes: Table computed at household and individual levels using survey information. Share of rooms with cement floors in 2000 is a self-declared retrospective variable that refers to the year 2000, while all the other variables are contemporaneous with the time of the survey. Standard errors clustered at census-block level shown in parentheses(136 clusters).
* Significantly different from 0 at 5 percent level.

None of the differences in means are significantly different from 0 at the 5 percent level, implying that the assumption the control and treatment groups are similarly distributed is valid.

1.4.2 Summary Statistics

Below are two tables of summary statistics from the household data, corresponding to the treatment and control groups respectively. Many of the variables in this table are indicator variables. Of interest, the S_cementfloor variables correspond to whether a household has a cement floor in a certain room. Additionally, we have S_cesds and S_pss which are used to measure the happiness of mothers using the CES-D Depression Scale and the PSS perceived stress scale. Other variables, both economic and health related, are summarized as well to give a cohesive overview of the data. While these statistics were not available in the original paper, we found that examining this table was beneficial to our understanding of the authors' methods and results.

Individual Data Control Group Summary Statistics

	Min	25%	Median	75%	Max	Mean	SD
Parasite Count	0	0	0	0	4	0.333	0.673
Diarrhea	0	0	0	0	1	0.142	0.349
Anemia	0	0	0	1	1	0.426	0.495
MCCDTS	0	1	6	16	91	13.354	18.952
PBDYPCT	3	12	23	45	97	30.656	24.864
Height-for-age Z-score	-4.72	-1.33	-0.6	0.1	2.96	-0.605	1.104
Weight-for-height Z-score	-3.89	-0.59	0.09	0.72	4.94	0.125	1.133
Respiratory Diseases	0	0	0	1	1	0.355	0.479
Skin Diseases	0	0	0	0	1	0.101	0.302
Other Diseases	0	0	0	0	1	0.041	0.198
Log Total Income of Mothers of Children 0-5 yrs	4.828	7.601	7.783	8.02	14.823	7.791	0.665
Log Total Income of Fathers of Children 0-5 yrs	3.219	7.818	8.071	8.302	14.503	8.121	0.592

Individual Data Treatment Group Summary Statistics

	Min	25%	Median	75%	Max	Mean	SD
Parasite Count	0	0	0	0	5	0.272	0.571
Diarrhea	0	0	0	0	1	0.124	0.33
Anemia	0	0	0	1	1	0.343	0.475
MCCDTS	0	3	9	22	96	17.391	20.988
PBDYPCT	3	12	25	50	97	33.132	25.954
Height-for-age Z-score	-4.78	-1.32	-0.59	0.1	2.99	-0.6	1.107
Weight-for-height Z-score	-4	-0.55	0.09	0.73	4.92	0.137	1.129
Respiratory Diseases	0	0	0	1	1	0.376	0.485
Skin Diseases	0	0	0	0	1	0.101	0.302
Other Diseases	0	0	0	0	1	0.046	0.21
Log Total Income of Mothers of Children 0-5 yrs	5.011	7.585	7.783	8.02	10.038	7.75	0.527
Log Total Income of Fathers of Children 0-5 yrs	4.2	7.88	8.071	8.294	14.5	8.106	0.601

Household Data Control Group Summary Statistics

	Min	25%	Median	75%	Max	Mean	SD
Overall	0	0.5	1	1	1	0.733	0.358
Kitchen	0	0	1	1	1	0.675	0.469
Dining	0	0	1	1	1	0.712	0.453
Bath	0	1	1	1	1	0.81	0.393
Bed	0	0	1	1	1	0.675	0.469
Satisfaction Floor	0	0	1	1	1	0.526	0.5
Satisfaction House	0	0	1	1	1	0.625	0.484
Satisfaction Life	0	0	1	1	1	0.616	0.487
CES-D Depression Score	0	12	19	24	56	18.598	9.513
Perceived Stress Score	0	12	17	20	39	16.454	6.965
Install Cement Floor	0	0	1	1	1	0.514	0.5
Install Sanitation Facilities	0	0	0	0	1	0.1	0.3
Restoration of Sanitation Facilities	0	0	0	0	1	0.045	0.207
Construction of Ceiling	0	0	0	0	1	0.151	0.358
Restoration of Walls	0	0	0	0	1	0.109	0.312
Improve Any	0	0	0	1	1	0.267	0.443
Log of Self-Reported Rental Value	3.401	5.521	5.991	6.215	11.385	5.921	0.760
Log of Self-Reported Sale Value	5.704	9.903	10.597	11.225	15.990	10.494	1.162
Total Consumption per Capita	0	447.5	623.520	822.333	29920.766	761.456	1299.443

Household Data Treatment Group Summary Statistics

	Min	25%	Median	75%	Max	Mean	SD
Overall	0	1	1	1	1	0.926	0.164
Kitchen	0	1	1	1	1	0.921	0.269
Dining	0	1	1	1	1	0.914	0.28
Bath	0	1	1	1	1	0.9	0.3
Bed	0	1	1	1	1	0.903	0.296
Satisfaction Floor	0	0	1	1	1	0.735	0.441
Satisfaction House	0	0	1	1	1	0.713	0.453
Satisfaction Life	0	0	1	1	1	0.717	0.451
CES-D Depression Score	0	11	16	22	47	16.414	8.441
Perceived Stress Score	0	10	15	19	36	14.891	6.504
Install Cement Floor	0	1	1	1	1	0.905	0.293
Install Sanitation Facilities	0	0	0	0	1	0.085	0.279
Restoration of Sanitation Facilities	0	0	0	0	1	0.04	0.197
Construction of Ceiling	0	0	0	0	1	0.188	0.391
Restoration of Walls	0	0	0	0	1	0.123	0.329
Improve Any	0	0	0	1	1	0.323	0.468
Log of Self-Reported Rental Value	2.996	5.704	5.991	6.215	11.385	5.949	0.793
Log of Self-Reported Sale Value	2.996	9.903	10.597	11.156	15.99	10.455	1.167
Total Consumption per Capita	0	492.196	641.25	877.75	35459.801	770.27	1137.67

In general the groups seem to be quite similar as asserted by the authors. While there are slight differences in some of the quantiles between the treatment and control groups, the overall distributions of each variable seem to be approximately the same. A few notable differences are that the CES-D and PSS scores for the treatment group seem to be lower overall as compared to the control group, which may have some influence on the results. However, the rest of the variables seem quite similarly distributed across the two groups.

2 Main Results

2.1 Assumptions

The authors of the paper made a few key assumptions, as they were fitting linear models. First, they assumed independence between the subjects being observed, namely that households in both the control and treatment groups had similar socioeconomic statuses, and a similar cultural and natural environment. By considering the multiple linear regression model,

$$y_{gi} = x'_{gi}\beta + u_{gi} \tag{1}$$

where observations belong to census-block level clusters $g = 1, \ldots, G$ and each observation in cluster is denoted as $i = 1, \ldots, n$. n is the number of observation per cluster.

The main results are calculated using clustered standard errors at the census level, meaning the standard assumptions associated with clusters are present here as

$$(X_g, y_g)_{g=1}^{G} \text{ i.i.d.} \tag{2}$$

This independence assumption assumes that observations per cluster are independent from the observations in all other clusters.

2.2 Main Result Replication

For each of the relevant variables, the authors fitted 3 linear models against variables of interest. Model 1 contains no controls, model 2 contains age, demographic, and health-habits controls, and model 3 contains age, demographic, health-habits, and public social programs controls. Tables 4-7 each fitted all three models against a group of dependent variables. Potentially due to some computational differences between R and Stata (the authors ran their analysis on Stata while we used R), some of the coefficients are off by a very small amount, usually around 0.001 or 0.002. However, these do not change their results, so we believe our replication is reliable.

The authors also check the significance of each coefficient at the 10, 5, and 1 percent levels. We only include the significance at the 5 percent level, however we do note if there are any significant results that the tables below omit. In keeping consistent with the table numbers from the original paper, we start from table 4 (for easy comparison with the orginal publication).

Table 4 contains regressions against variables representing the cement floor coverage level in houses. These regressions indicate that Piso Firme did indeed have a strong effect on increasing the share of cement floors in houses. Overall, Piso Firme increased the share of cement floors around 28% on average across all rooms, with sleeping areas and dining areas seeing the largest increase. All of the coefficients were found to be significant at the 5% level.

Table 4: Regressions of Cement Floor Coverage Measures on Program Dummy

		Model 1			Model 2			Model 3		
Variable	**Control Mean/SD**	**Coef1**	**RobustSE1**	**Fraction1**	**Coef2**	**RobustSE2**	**Fraction2**	**Coef3**	**RobustSE3**	**Fraction3**
Overall	0.728 (0.363)	0.201*	0.020	27.610	0.207*	0.019	28.434	0.210*	0.019	28.846
Kitchen	0.671 (0.470)	0.255*	0.025	38.003	0.260*	0.022	38.748	0.265*	0.023	39.493
Dining	0.709 (0.455)	0.210*	0.025	29.619	0.217*	0.025	30.606	0.221*	0.025	31.171
Bath	0.803 (0.398)	0.101*	0.022	12.578	0.111*	0.018	13.823	0.114*	0.018	14.197
Bed	0.668 (0.471)	0.239*	0.020	35.778	0.246*	0.020	36.826	0.246*	0.020	36.826

Notes: Model 1: no controls; Model 2: age, demographic, and health-habits 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 5 percent level.

Table 5 contains regressions against variables representing children's health measures. These are specifically examined for children from 0-5 years of age, and include parasite count, presence of diarrhea and anemia, as well as test scores measuring literacy and cognative development. The authors found these to be significant at the 5% level, with a few of the coefficients such as those for anemia being significant at the 1% level. Overall, we were able to replicate these findings and find the same significant coefficients.

Table 5: Regressions of Children's Health Measures on Program Dummy

		Model 1			Model 2			Model 3		
Variable	**Control Mean/SD**	**Coef1**	**RobustSE1**	**Fraction1**	**Coef2**	**RobustSE2**	**Fraction2**	**Coef3**	**RobustSE3**	**Fraction3**
Parasite Count	0.333 (0.673)	-0.061*	0.032	-18.318	-0.064*	0.030	-19.219	-0.063*	0.030	-18.919
Diarrhea	0.142 (0.349)	-0.018	0.009	-12.676	-0.019*	0.009	-13.380	-0.018	0.009	-12.676
Anemia	0.426 (0.495)	-0.083*	0.028	-19.484	-0.074*	0.027	-17.371	-0.077*	0.027	-18.075
MCCDTS	13.354 (18.952)	4.037*	1.642	30.231	4.598*	1.547	34.432	4.632*	1.534	34.686
PBDYPCT	30.656 (24.864)	2.476	1.681	8.077	2.715*	1.526	8.856	2.536*	1.526	8.272
Height-for-age Z-score	-0.605 (1.104)	0.005	0.043	-0.826	-0.003	0.040	0.496	-0.001	0.041	0.165
Weight-for-height Z-score	0.125 (1.133)	0.012	0.034	9.600	0.004	0.035	3.200	0.000	0.036	0.000

Notes: Variable abbreviations: MCCDTS - MacArthur Communicative Development Test Score, PBDYPCT - Picture Peabody Vocabulary Test percentile score. Model 1: no controls; Model 2: age, demographic, and health-habits 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 5 percent level.

Table 6 contains variables pertaining to maternal happiness and mental health. These variables include satisfcation with the floor quality, house quality, and life quality. They also include two scores on the CES-D depression scale as well as the PSS (perceived stress scale). The authors found all of these coefficients to be statistically significant at the 5% level. This tells us that the Piso Firme did in fact improve maternal happiness and mental health as measured by these variables.

Table 6: Regressions of Satisfaction and Maternal Mental Health Measures on Program Dummy

		Model 1			Model 2			Model 3		
Variable	**Control Mean/SD**	**Coef1**	**RobustSE1**	**Fraction1**	**Coef2**	**RobustSE2**	**Fraction2**	**Coef3**	**RobustSE3**	**Fraction3**
Satisfaction Floor	0.511 (0.500)	0.221*	0.023	43.249	0.225*	0.024	44.031	0.224*	0.025	43.836
Satisfaction House	0.605 (0.489)	0.095*	0.021	15.702	0.089*	0.021	14.711	0.086*	0.022	14.215
Satisfaction Life	0.601 (0.490)	0.111*	0.022	18.469	0.110*	0.021	18.303	0.111*	0.022	18.469
CES-D Depression Scale	18.532 (9.402)	-2.207*	0.614	-11.909	-2.361*	0.567	-12.740	-2.342*	0.562	-12.638
Perceived Stress Scale	16.514 (6.914)	-1.721*	0.426	-10.421	-1.763*	0.396	-10.676	-1.753*	0.398	-10.615

Notes: Variable abbreviations: MCCDTS - MacArthur Communicative Development Test Score, PBDYPCT - Picture Peabody Vocabulary Test percentile score. Model 1: no controls; Model 2: age, demographic, and health-habits 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 5 percent level.

2.3 Assumption Critique

The authors utilized the cluster standardized errors in their analysis. This means that they assumed that all of the clusters were independent of one another, and that these clusters each contribute to the uncertainty independently. The authors do their own critique of this assumption in their robustness checks, by computing a between-cluster Moran's I test statistic between each cluster. They fail to reject the null hypothesis of zero spacial autocorrelation between the clusters.

The authors also choose to cluster at the census block level. While this may be the most logical choice, it is possible that clustering at either the smaller neighborhood level or the larger state level could lead to different results in this context. Due to the clustering being done at the census block level, it is possible that these clusters are large enough that we would see different effects within the same cluster.

3 Robustness Check

One robustness implemented in the paper was a falsification test to see if there were other possible reasons besides Piso Firme for the results observed. They conducted this test by seeing if there were any differences between the rates of other diseases between the treatment and control groups, thus implying that there could be some other factor contributing instead of Piso Firme. The authors fitted 3 models similar to the main result, only this time looking at these other diseases in relationship to covariates. The results of this test (table 7 in the paper) are shown below:

Table 7: Robustness Checks

		Model 1			Model 2			Model 3		
Variable	**Control Mean/SD**	**Coef1**	**RobustSE1**	**Fraction1**	**Coef2**	**RobustSE2**	**Fraction2**	**Coef3**	**RobustSE3**	**Fraction3**
Respiratory Diseases	0.355 (0.479)	0.021	0.019	5.915	0.020	0.018	5.634	0.018	0.019	5.070
Skin Diseases	0.101 (0.302)	0.001	0.003	0.002	-0.001	0.011	-0.990	-0.002	0.011	-1.980
Other Diseases	0.041 (0.198)	0.005	0.009	12.195	0.005	0.009	12.195	0.005	0.009	12.195
Inst. of Cement Floor	0.530 (0.499)	0.376*	0.028	70.943	0.375*	0.028	70.755	0.376*	0.028	70.943
Const. of Sanitation Facilities	0.101 (0.302)	-0.017	0.015	-16.832	-0.018	0.015	-17.822	-0.017	0.015	-16.832
Rest. of Sanitation Facilities	0.045 (0.206)	-0.001	0.013	-2.222	-0.001	0.013	-2.222	-0.002	0.012	-4.444
Const. of Ceiling	0.159 (0.366)	0.028	0.024	17.610	0.021	0.024	13.208	0.018	0.023	11.321
Rest. of Walls	0.111 (0.314)	0.012	0.017	10.811	0.012	0.015	10.811	0.014	0.016	12.613
House Expansion	0.277 (0.448)	0.045	0.031	16.245	0.037	0.031	13.357	0.038	0.030	13.718
Log of Rent Value	5.918 (0.740)	0.035	0.040	0.591	0.053	0.032	0.896	0.056	0.031	0.946
Log of Sell Value	10.491 (1.168)	-0.043	0.100	-0.410	-0.017	0.083	-0.162	-0.014	0.080	-0.133
Log of Income of Mothers	7.791 (0.665)	-0.042	0.064	-0.539	-0.039	0.065	-0.501	-0.045	0.065	-0.578
Log of Income of Fathers	8.121 (0.592)	-0.015	0.027	-0.185	-0.005	0.026	-0.062	0.009	0.027	0.111
Total Consumption per Capita	753.733 (1219.488)	4.272	44.197	0.567	12.992	77.130	1.724	16.146	256.246	2.142

Variable abbreviations: MCCDTS - MacArthur Communicative Development Test Score, PBDYPCT - Picture Peabody Vocabulary Test percentile score. Model 1: no controls; Model 2: age, demographic, and health-habits 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 5 percent level.

As we can see from the table, only the coefficient for "Installation of Cement Floor" was statistically different from 0 at the 10, 5, and 1 percent levels. This robustness check confirms that it was indeed the installation of these cement floors through the Piso Firme program that led to the results observed, and not some other factor not included in the model.

4 Re-analysis

4.1 Leverage Score

We re-analyzed the results of this paper utilizing a few methods from class. First, we computed the leverage scores of the models and identify any highly influential points. Looking at the leverage score plots, we find that for model 1 (no control), all of the points have similar leverage scores. However, for models 2 and 3, we find that there a few points with significantly large leverage scores. We summarise some observations about these points in table 8 below.

There are 2 points coming from the control group, and one from the treatment group. The most influential point is the one from the treatment group (index 2128), with a leverage score of 1.0 in model 3. For reference, the average leverage score was around 0.009, which means this point was highly influential compared to every other point. Looking closer at the observation, it seems like many of the reported values differ greatly from the average values observed in the summary statistics table. For example, they reported 0 household members, but still reported a head of household age and spouse age. There is also no reported income per capita, assets per capita, or consumption per capita.

The other two points from the control group are less extreme, but there are still clear deviations from the "average" household. Repeating the analysis without these three points changes the results significantly, with the cluster-robust errors increasing by a magnitude of around 10x across all 3 models. However, points 660 and 1262 seems reasonable, it is only 2128 that seems like a large outlier.

Running the analysis without point 2128 has a significant effect on model 3. We found that all of the robust standard errors for model 3 increased over tenfold. This is a significant finding, as it is clear that this point is an outlier with extreme values. Given the scope of this project, we did not do further analysis, however we believe it is worth looking into repeating the analysis after examining whether to include outliers such as this one.

Table 8: High Leverage Points Data Record

Variable	660	1262	2128
dpisofirme	0	0	1
Num of Household Members	5	3	0
Head of Household Age	30	41	18
Spouse Age	28	0	17
Head of Household Educ.	6	6	0
Spouse Educ.	6	NA	8
Num of Rooms	2	1	1
Water Connection (=1)	0	0	1
Water Connection in House (=1)	0	0	0
Electricity (=1)	0	0	1
Share of Cement Floors 2000	0	0	0
Animals on Land (=1)	0	1	1
Animals Inside (=1)	0	0	0
Garbage Collection (=1)	1	1	1
Num Washed Hands	0	4	3
Income Per Capita	486.400	810.667	0
Assets Per Capita	20635.980	35897.527	NA
Consumption Per Capita	259.800	829.533	NA
Cash Transfers from Gov	0	175	NA
Go Milk Program (=1)	1	1	0
Gov Food Program (=1)	0	1	0
Satisfied with Floor (=1)	0	1	1
Satisfied with House (=1)	0	1	1
Satisfied with Life Quality (=1)	0	1	1
CES-D Score	35	30	27
PSS Score	20	18	9
Installation of Cement Floor	0	1	1
Installation of Sanitation Fac. (=1)	0	0	0
Restoration of Sanitation Fac. (=1)	0	0	0
Construction of Ceiling (=1)	0	0	0
Restoration of Walls (=1)	0	0	0
Any Improvement (=1)	0	0	0
Log of Self-Reported Rent Value	5.704	5.298	5.298
Log of Self-Reported Sell Value	8.517	10.463	7.601

4.2 Leave-One-Out Cross-Validation

We also performed leave-one-out cross-validation in order to examine the linear models from table 4-6. Leave-one-out Cross-Validation provides a much less biased measure of test MSE and tends not to overestimate the test MSE compared to using a single test set.

From LOOCV, our re-analysis focus on the performance on models. The RMSE is the root mean squared error, measuring the average difference between predictions and observations.

R^2 is a score from 0-1 to measure the correlation between predictions and observations. MAE is the average absolute difference between predictions and observation. Therefore, we expect the RMSE and MAE to be low and R^2 to be high for good modeling.

As a result, Table 9 collects the output of RMSE, R^2 and MAE for each model from Table 4-6 by leave one out cross-validation.

Based on RMSE and MAE, Model 2 and Model 3 outperformed Model 1 in general. In terms that, Single predictor is not sufficient to accurately predict the outcomes. While low R^2 values are common in social science data, the R^2 across all our models is around or below 0.2, indicating that the correlation between predicted values and observations is not large for overall models. This could imply some non-linearity in the data.

By analyzing the leave-one-out cross-validation results, we suspect that our variables of interests do not have a strong linear relationship with our models. Multivariate linear regression might not be the most ideal methods to regress the variables of interest, and we would recommend exploring non-linear methods as well.

Table 9: Leave One Out Cross Validation

	Model1			**Model2**			**Model3**		
Variable	**RMSE**	**Rsquared**	**MAE**	**RMSE**	**Rsquared**	**MAE**	**RMSE**	**Rsquared**	**MAE**
Overall Floor	0.281	0.112	0.212	0.266	0.174	0.197	0.266	0.174	0.197
Kitchen Floor	0.381	0.099	0.289	0.368	0.147	0.281	0.368	0.149	0.282
Dining Floor	0.375	0.071	0.282	0.364	0.102	0.271	0.364	0.104	0.271
Bathroom Floor	0.351	0.019	0.246	0.335	0.080	0.230	0.335	0.080	0.231
Bedroom Floor	0.392	0.084	0.307	0.380	0.118	0.293	0.381	0.115	0.293
Parasite Count	0.625	0.001	0.464	0.620	0.029	0.443	0.622	0.026	0.444
Diarrhea	0.340	0.000	0.231	0.339	0.019	0.229	0.339	0.016	0.229
Anemia	0.485	0.006	0.471	0.477	0.039	0.453	0.478	0.039	0.452
MCCDTS	20.025	0.005	14.501	17.078	0.217	11.717	17.286	0.203	11.947
PBDYPCT	25.416	0.001	20.837	24.934	0.045	20.401	24.897	0.051	20.312
Height for Age Z-Score	1.106	0.052	0.866	1.102	0.007	0.862	1.100	0.009	0.860
Weight for Height Z-Score	1.131	0.007	0.842	1.135	0.004	0.843	1.132	0.007	0.842
Satisfaction Floor	0.473	0.051	0.446	0.473	0.052	0.443	0.473	0.050	0.443
Satisfaction House	0.474	0.009	0.449	0.474	0.013	0.445	0.474	0.012	0.445
Satisfaction Life	0.472	0.012	0.445	0.468	0.021	0.434	0.469	0.019	0.434
CES-D Depression Scale	8.892	0.014	7.046	8.716	0.043	6.874	8.724	0.042	6.879
Perceived Stress Scale	6.679	0.015	5.316	6.649	0.021	5.266	6.651	0.020	5.268

5 Conclusion

The paper "Housing, Health, and Happiness" by Matias Cattaneo et al. sought to explore the relationship between the presence of cement floors and measures of health and happiness. Using linear models, they were able to show that there is indeed a statistically significant improvement in health and happiness through the installation of cement floors through the Piso Firme program. We were able to replicate their main results, as well as one of the robustness checks. A few of our coefficients ended up being slightly different from the ones obtained in the original analysis, however we believe this to be due to computational differences between R and Stat.

Furthermore, We re-analyzed their methods utilizing leverage scores and LOOE on outliers and model evaluation. We concluded that this analysis, while robust to various specifications, includes potential outliers that could be affecting the findings. We found that a specific household had an extremely high leverage score compared to the other households, and redoing the analysis without this point caused the robust standard errors to increase over tenfold. While we were unable to do more in-depth analysis, we believe that this could potentially be an oversight and should be looked into further.

From the LOOCV, we found that the model 2 and model 3 outperformed model 1 which indicates that single predictor of control/treatment group indicator is not sufficient to predict our outcomes. Furthermore, since R^2 are low for models in general, the data might not have strong linear relationships and thus alternative models should also be considered.

References

[1] Matias D Cattaneo, Sebastian Galiani, Paul J Gertler, Sebastian Martinez, and Rocio Titiunik. Housing, health, and happiness. *American Economic Journal: Economic Policy*, 1(1):75–105, 2009.

230 Project EDA

Daniel Huang, Yiyun Gong

4/6/2022

```r
#Read in data
library(haven)
individual <- read_dta("Data/PisoFirme_AEJPol-20070024_individual.dta")
household <- read_dta("Data/PisoFirme_AEJPol-20070024_household.dta")
```

Sample Sizes:

```r
#Returns sample size for one column
my_func <- function(x) {
  return(sum(!is.na(x)))
}

#Household
sample_sizes <- household %>%
  filter(!is.na(idcluster)) %>%
  group_by(dpisofirme) %>%
  summarise_all(my_func)

h_ss <- as.data.frame(t(sample_sizes))
h_ss

write.csv(h_ss, file="sample_sizes_h.csv")
```

```r
#Individual
sample_sizes2 <- individual %>%
  filter(!is.na(idcluster)) %>%
  group_by(dpisofirme) %>%
  summarise_all(my_func)

i_ss <- as.data.frame(t(sample_sizes2))
i_ss

write.csv(i_ss, file="sample_sizes_i.csv")
```

Summary Statistics for both groups in Household:

```r
my_func <- function(x) {
  return(c(round(fivenum(x), 3), round(mean(x, na.rm = T), 3), round(sd(x, na.rm = T), 3)))
}
household_num <- household %>%
  select(c("dpisofirme", "S_shcementfloor", "S_cementfloorkit", "S_cementfloordin", "S_cementfloorbat",

household0 <- household_num %>%
  drop_na() %>%
```

```r
  filter(dpisofirme == 0) %>%
  select(-dpisofirme)

household1 <- household_num %>%
  drop_na() %>%
  filter(dpisofirme == 1) %>%
  select(-dpisofirme)

result0 <- apply(household0, 2, my_func)
df0 <- as.data.frame(t(result0))
colnames(df0) <- c("Min", "25%", "Median", "75%", "Max", "Mean", "Standard Deviation")

result1 <- apply(household1, 2, my_func)
df1 <- as.data.frame(t(result1))
colnames(df1) <- c("Min", "25%", "Median", "75%", "Max", "Mean", "Standard Deviation")

#write.table(df0, file = "df0.txt", sep = ",", quote = FALSE, row.names = T)
#write.table(df1, file = "df1.txt", sep = ",", quote = FALSE, row.names = T)

#HH control
write.table(df0, "clipboard-16384", sep = "\t", row.names = T, quote = F)
#HH Treat
#write.table(df1, "clipboard-16384", sep = "\t", row.names = T, quote = F)
```

Summary statistics for both groups in Individual:

```r
individual_num <- individual %>%
  select("dpisofirme", "S_parcount", "S_diarrhea", "S_anemia","S_mccdts",
"S_pbdypct", "S_haz","S_whz","S_respira","S_skin", "S_otherdis",
"S_malincom", "S_palincom")

ind0 <- individual_num %>%
  filter(!is.na(dpisofirme)) %>%
  filter(dpisofirme == "0") %>%
  select(-dpisofirme)

ind1 <- individual_num %>%
  filter(!is.na(dpisofirme)) %>%
  filter(dpisofirme == "1") %>%
  select(-dpisofirme)

res0 <- apply(ind0, 2, my_func)
summ_ind_0 <- as.data.frame(t(res0))
colnames(summ_ind_0) <- c("Min", "25%", "Median", "75%", "Max", "Mean", "Standard Deviation")

res1 <- apply(ind1, 2, my_func)
summ_ind_1 <- as.data.frame(t(res1))
colnames(summ_ind_1) <- c("Min", "25%", "Median", "75%", "Max", "Mean", "Standard Deviation")

#Ind control
#write.table(summ_ind_0, "clipboard-16384", sep = "\t", row.names = T, quote = F)

#Ind Treat
write.table(summ_ind_1, "clipboard-16384", sep = "\t", row.names = T, quote = F)
```

Table 3 Generator

```
# read in data
household <- read_dta("household.dta")
household <- household[!is.na(household$idcluster),]

# Split into treatment/Control group
household_ctr = household[which(household$dpisofirme == 0),]
household_trt = household[which(household$dpisofirme == 1),]

HH_survey = c("S_HHpeople","S_headage","S_spouseage",
              "S_headeduc","S_spouseeduc","S_rooms",
              "S_waterland", "S_waterhouse", "S_electricity",
              "S_cementfloor2000", "S_hasanimals",
              "S_animalsinside", "S_garbage", "S_washhands",
              "S_incomepc", "S_assetspc", "S_shpeoplework",
              "S_microenter", "S_hrsworkedpc", "S_consumptionpc",
              "S_cashtransfers", "S_milkprogram", "S_foodprogram")

table3 = data.frame(Variable = character(),
                          obs_trt = numeric(),
                          Mean_trt = numeric(),
                          obs_ctr = numeric(),
                          Mean_ctr = numeric(),
                          Mean_diff = numeric()
                           )
for(i in 1:length(HH_survey)){
  table3[i,1] = HH_survey[i]
  table3[i,2] = length(which(!is.na(household_trt[[HH_survey[i]]])))
  table3[i,3] = round(mean(household_trt[[HH_survey[i]]], na.rm=TRUE), 3)
  table3[i,4] = length(which(!is.na(household_ctr[[HH_survey[i]]])))
  table3[i,5] = round(mean(household_ctr[[HH_survey[i]]], na.rm=TRUE), 3)
}
table3$Mean_diff = table3$Mean_trt - table3$Mean_ctr
table3

CH_survey=c("S_age", "S_gender", "S_childma", "S_childmaage",
            "S_childmaeduc", "S_childpa", "S_childpaage",
            "S_childpaeduc")

# read in data
individual <- read_dta("individual.dta")
individual <- individual[which(individual$S_age <= 5),]
individual <- individual[!is.na(individual$idcluster),]

# Split into treatment/Control group
individual_ctr = individual[which(individual$dpisofirme == 0),]
individual_trt = individual[which(individual$dpisofirme == 1),]
```

```r
table4 = data.frame(Variable = character(),
                          obs_trt = numeric(),
                          Mean_trt = numeric(),
                          obs_ctr = numeric(),
                          Mean_ctr = numeric(),
                          Mean_diff = numeric()
                           )
for(i in 1:length(CH_survey)){
  table4[i,1] = CH_survey[i]
  table4[i,2] = length(which(!is.na(individual_trt[[CH_survey[i]]])))
  table4[i,3] = round(mean(individual_trt[[CH_survey[i]]], na.rm=TRUE), 3)
  table4[i,4] = length(which(!is.na(individual_ctr[[CH_survey[i]]])))
  table4[i,5] = round(mean(individual_ctr[[CH_survey[i]]], na.rm=TRUE), 3)
}
table4$Mean_diff = table4$Mean_trt - table4$Mean_ctr
table4

table3 = rbind(table3, table4)

write.csv(table3, "table3.csv")
```

Individual Data Modeling

```
# read in data
individual <- read_dta("individual.dta")
# Split into treatment/Control group
individual_ctr = individual[which(individual$dpisofirme == 0),]
individual_trt = individual[which(individual$dpisofirme == 1),]
```

Variable Definition

```
# household variables
demog1 = c("S_HHpeople", "S_headage", "S_spouseage",
           "S_headeduc", "S_spouseeduc")
demog2 = c("S_dem1", "S_dem2", "S_dem3", "S_dem4",
           "S_dem5", "S_dem6", "S_dem7", "S_dem8")
health = c("S_waterland", "S_waterhouse", "S_electricity",
           "S_hasanimals", "S_animalsinside", "S_garbage",
           "S_washhands")
econ = c("S_incomepc", "S_assetspc")
# remove cashtransfers from social
social = c("S_milkprogram", "S_foodprogram", "S_seguropopular")
floor = c("S_shcementfloor", "S_cementfloorkit",
          "S_cementfloordin", "S_cementfloorbat",
          "S_cementfloorbed")
satis = c("S_satisfloor", "S_satishouse", "S_satislife",
          "S_cesds", "S_pss")
robust = c("S_instcement", "S_instsanita", "S_restsanita",
           "S_constceili", "S_restowalls", "S_improveany",
           "S_logrent", "S_logsell", "S_consumptionpc")
cash = c("S_cashtransfers")
```

```
# individual variables
CH_survey = c("S_age","S_gender", "S_childma", "S_childmaage",
              "S_childmaeduc", "S_childpa", "S_childpaage",
              "S_childpaeduc")
CH_demog = c("S_HHpeople", "S_rooms", "S_age", "S_gender",
             "S_childma", "S_childmaage", "S_childmaeduc",
             "S_childpa", "S_childpaage", "S_childpaeduc")
CH_health = c("S_parcount", "S_diarrhea", "S_anemia", "S_mccdts",
              "S_pbdypct", "S_haz", "S_whz")
CH_robust = c("S_respira", "S_skin", "S_otherdis")
PA_robust = c("S_malincom", "S_palincom")
```

```
# add missing value indicator column names in demog1, demog2, health, econ, cashtransfer
demog1_miss = paste0(demog1, "_Miss")
```

```r
demog2_miss = paste0(demog1, "_Miss")
health_miss = paste0(health, "_Miss")
econ_miss = paste0(econ, "_Miss")
cash_miss = paste0(cash, "_Miss")
CH_demog_miss = paste0(CH_demog, "_Miss")
```

Missing Value Imputations

```r
# individual
# Columns: demog1, demog2, health and econ
mis_ls = c(CH_demog, health, econ)
mis_actual = c()
for (i in 1:length(mis_ls)){
  # Add corresponding dummy variable columns to indicate missing
  mis_name = paste0(mis_ls[i], '_Miss')
  individual[[mis_name]] = ifelse(is.na(individual[[mis_ls[i]]]), 1, 0)
  # Impute Missing Values of columns with 0
  individual[[mis_ls[i]]][is.na(individual[[mis_ls[i]]])] <- 0
  if(sum(individual[[mis_name]]) != 0){
    mis_actual = c(mis_actual, mis_name)
  }
}
```

Model Definition

```r
#household
mod1 = '~dpisofirme'

mod2 = paste0(c(CH_demog, health), collapse = '+')
mod2 = paste0(mod1, "+", mod2)

mod3 = paste0(c(CH_demog, health, cash, social, econ), collapse = '+')
mod3 = paste0(mod1, "+", mod3)
```

Table 5 Children's Health (Individual)

```r
# Control group mean & SD for floor list
chealth_reg_tbl = data.frame(Variable = character(),
                             Mean_Ctr = numeric(),
                             SD_Ctr = numeric(),
                             M1_Coef = numeric(),
                             M1_RobustSE = numeric(),
                             M1_fraction = numeric(),
                             M2_Coef = numeric(),
                             M2_RobustSE = numeric(),
                             M2_fraction = numeric(),
                             M3_Coef = numeric(),
```

```
                           M3_RobustSE = numeric(),
                           M3_fraction = numeric()
                           )
for (i in 1:length(CH_health)){
  chealth_reg_tbl[i,1] = CH_health[i]
  print(CH_health[i])
  chealth_reg_tbl[i,2] = round(mean(individual_ctr[[CH_health[i]]], na.rm=TRUE), 3)
  print(round(mean(individual_ctr[[CH_health[i]]]), 3))
  chealth_reg_tbl[i,3] = round(sd(individual_ctr[[CH_health[i]]], na.rm=TRUE), 3)
}
chealth_reg_tbl
```

```
# Model 1
# Estimated Coefficient for CH_health
for (i in 1:length(CH_health)){
  reg = paste0(CH_health[i], mod1)
  chealth_reg_tbl[i,4] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = individual)
  chealth_reg_tbl[i,5] = round(summary(Pten.gee)$coef[2,4], 3)
}

chealth_reg_tbl$M1_fraction = round((chealth_reg_tbl$M1_Coef*100)/chealth_reg_tbl$Mean_Ctr, 3)
chealth_reg_tbl

# Model 2
# Estimated Coefficient for CH_health
for (i in 1:length(CH_health)){
  reg = paste0(CH_health[i], mod2)
  chealth_reg_tbl[i,7] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = individual)
  chealth_reg_tbl[i,8] = round(summary(Pten.gee)$coef[2,4], 3)
}

chealth_reg_tbl$M2_fraction = round((chealth_reg_tbl$M2_Coef*100)/chealth_reg_tbl$Mean_Ctr, 3)
chealth_reg_tbl

# Model 3
# Estimated Coefficient for CH_health
for (i in 1:length(CH_health)){
  reg = paste0(CH_health[i], mod3)
  chealth_reg_tbl[i,10] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
```

```
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = individual)
  chealth_reg_tbl[i,11] = round(summary(Pten.gee)$coef[2,4], 3)
}

chealth_reg_tbl$M3_fraction = round((chealth_reg_tbl$M3_Coef*100)/chealth_reg_tbl$Mean_Ctr, 3)
chealth_reg_tbl
```

Table 7 Robustness Checks (Individual dataset)

```
# Control group mean & SD for Robustness Checks list
Rob_reg_tbl = data.frame(Variable = character(),
                           Mean_Ctr = numeric(),
                           SD_Ctr = numeric(),
                           M1_Coef = numeric(),
                           M1_RobustSE = numeric(),
                           M1_fraction = numeric(),
                           M2_Coef = numeric(),
                           M2_RobustSE = numeric(),
                           M2_fraction = numeric(),
                           M3_Coef = numeric(),
                           M3_RobustSE = numeric(),
                           M3_fraction = numeric()
                           )
for (i in 1:length(CH_robust)){
  Rob_reg_tbl[i,1] = CH_robust[i]
  Rob_reg_tbl[i,2] = round(mean(individual_ctr[[CH_robust[i]]], na.rm=TRUE), 3)
  Rob_reg_tbl[i,3] = round(sd(individual_ctr[[CH_robust[i]]], na.rm=TRUE), 3)
}

# Model 1
# Estimated Coefficient for CH_robust
for (i in 1:length(CH_robust)){
  reg = paste0(CH_robust[i], mod1)
  Rob_reg_tbl[i,4] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # CH_robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = individual)
  Rob_reg_tbl[i,5] = round(summary(Pten.gee)$coef[2,4], 3)
}

Rob_reg_tbl$M1_fraction = round((Rob_reg_tbl$M1_Coef*100)/Rob_reg_tbl$Mean_Ctr, 3)

# Model 2
# Estimated Coefficient for floor
for (i in 1:length(CH_robust)){
```

```
  reg = paste0(CH_robust[i], mod2)
  Rob_reg_tbl[i,7] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # CH_robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = individual)
  Rob_reg_tbl[i,8] = round(summary(Pten.gee)$coef[2,4], 3)
}

Rob_reg_tbl$M2_fraction = round((Rob_reg_tbl$M2_Coef*100)/Rob_reg_tbl$Mean_Ctr, 3)

# Model 3
# Estimated Coefficient for floor
for (i in 1:length(CH_robust)){
  reg = paste0(CH_robust[i], mod3)
  Rob_reg_tbl[i,10] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # CH_robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = individual)
  Rob_reg_tbl[i,11] = round(summary(Pten.gee)$coef[2,4], 3)
}

Rob_reg_tbl$M3_fraction = round((Rob_reg_tbl$M3_Coef*100)/Rob_reg_tbl$Mean_Ctr, 3)
Rob_reg_tbl
```

Table 7 PA Robustness Checks (Individual dataset)

```
# Control group mean & SD for Robustness Checks list
PARob_reg_tbl = data.frame(Variable = character(),
                           Mean_Ctr = numeric(),
                           SD_Ctr = numeric(),
                           M1_Coef = numeric(),
                           M1_RobustSE = numeric(),
                           M1_fraction = numeric(),
                           M2_Coef = numeric(),
                           M2_RobustSE = numeric(),
                           M2_fraction = numeric(),
                           M3_Coef = numeric(),
                           M3_RobustSE = numeric(),
                           M3_fraction = numeric()
                           )
for (i in 1:length(PA_robust)){
  PARob_reg_tbl[i,1] = PA_robust[i]
  PARob_reg_tbl[i,2] = round(mean(individual_ctr[[PA_robust[i]]], na.rm=TRUE), 3)
  PARob_reg_tbl[i,3] = round(sd(individual_ctr[[PA_robust[i]]], na.rm=TRUE), 3)
}
```

```r
# Model 1
# Estimated Coefficient for PA_robust
for (i in 1:length(PA_robust)){
  reg = paste0(PA_robust[i], mod1)
  PARob_reg_tbl[i,4] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # PA_robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    na.action = na.omit,
    data = individual)
  PARob_reg_tbl[i,5] = round(summary(Pten.gee)$coef[2,4], 3)
}

PARob_reg_tbl$M1_fraction = round((PARob_reg_tbl$M1_Coef*100)/PARob_reg_tbl$Mean_Ctr, 3)
PARob_reg_tbl

# Model 2
# Estimated Coefficient for floor
for (i in 1:length(PA_robust)){
  reg = paste0(PA_robust[i], mod2)
  PARob_reg_tbl[i,7] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # PA_robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    na.action = na.omit,
    data = individual)
  PARob_reg_tbl[i,8] = round(summary(Pten.gee)$coef[2,4], 3)
}

PARob_reg_tbl$M2_fraction = round((PARob_reg_tbl$M2_Coef*100)/PARob_reg_tbl$Mean_Ctr, 3)
PARob_reg_tbl

# Model 3
# Estimated Coefficient for floor
for (i in 1:length(PA_robust)){
  reg = paste0(PA_robust[i], mod3)
  PARob_reg_tbl[i,10] = round(summary(lm(reg, data = individual))$coefficients[2, 1], 3)
  # PA_robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    na.action = na.omit,
    data = individual)
  PARob_reg_tbl[i,11] = round(summary(Pten.gee)$coef[2,4], 3)
}

PARob_reg_tbl$M3_fraction = round((PARob_reg_tbl$M3_Coef*100)/PARob_reg_tbl$Mean_Ctr, 3)
```

```
PARob_reg_tbl
```

LOOE CV

Children's health

```
# LOOCV Error
library(caret)
#specify the cross-validation method
ctrl <- trainControl(method = "LOOCV")
```

```
# CH_health mod1
for(i in 1:length(CH_health)){
  modd = as.formula(paste0(CH_health[i], mod1))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = individual, method = "lm", trControl = ctrl, na.action = 'na.omit')
  print(paste(CH_health[i], "mod1"))
  print(model)
}
```

```
# CH_health mod2
for(i in 1:length(CH_health)){
  modd = as.formula(paste0(CH_health[i], mod2))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = individual, method = "lm", trControl = ctrl, na.action = 'na.omit')
  print(paste(CH_health[i], "mod2"))
  print(model)
}
```

```
# CH_health mod2
for(i in 1:length(CH_health)){
  modd = as.formula(paste0(CH_health[i], mod3))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = individual, method = "lm", trControl = ctrl, na.action = 'na.omit')
  print(paste(CH_health[i], "mod3"))
  print(model)
}
```

Household Data Modeling

```r
# read in data
individual <- read_dta("individual.dta")
household <- read_dta("household.dta")
# Split into treatment/Control group
household_ctr = household[which(household$dpisofirme == 0),]
household_trt = household[which(household$dpisofirme == 1),]
```

Variable Definition

```r
# household variables
demog1 = c("S_HHpeople", "S_headage", "S_spouseage",
           "S_headeduc", "S_spouseeduc")
demog2 = c("S_dem1", "S_dem2", "S_dem3", "S_dem4",
           "S_dem5", "S_dem6", "S_dem7", "S_dem8")
health = c("S_waterland", "S_waterhouse", "S_electricity",
           "S_hasanimals", "S_animalsinside", "S_garbage",
           "S_washhands")
econ = c("S_incomepc", "S_assetspc")
# remove cashtransfers from social
social = c("S_milkprogram", "S_foodprogram", "S_seguropopular")
floor = c("S_shcementfloor", "S_cementfloorkit",
          "S_cementfloordin", "S_cementfloorbat",
          "S_cementfloorbed")
satis = c("S_satisfloor", "S_satishouse", "S_satislife",
          "S_cesds", "S_pss")
robust = c("S_instcement", "S_instsanita", "S_restsanita",
           "S_constceili", "S_restowalls", "S_improveany",
           "S_logrent", "S_logsell", "S_consumptionpc")
cash = c("S_cashtransfers")
```

```r
# add missing value indicator column names in demog1, demog2, health, econ, cashtransfer
demog1_miss = paste0(demog1, "_Miss")
demog2_miss = paste0(demog1, "_Miss")
health_miss = paste0(health, "_Miss")
econ_miss = paste0(econ, "_Miss")
cash_miss = paste0(cash, "_Miss")
```

Missing Value Imputations

```r
# Household
# Columns: demog1, demog2, health and econ
```

```
mis_ls = c(demog1, demog2, health, econ, cash)
mis_actual = c()
for (i in 1:length(mis_ls)){
  # Add corresponding dummy variable columns to indicate missing
  mis_name = paste0(mis_ls[i], '_Miss')
  household[[mis_name]] = ifelse(is.na(household[[mis_ls[i]]]), 1, 0)
  # Impute Missing Values of columns with 0
  household[[mis_ls[i]]][is.na(household[[mis_ls[i]]])] <- 0
  if(sum(household[[mis_name]]) != 0){
    mis_actual = c(mis_actual, mis_name)
  }
}
```

Model Definition

```
#household
mod1 = '~dpisofirme'

mod2 = paste0(c(demog1, demog2, health), collapse = '+')
mod2 = paste0(mod1, "+", mod2)

mod3 = paste0(c(demog1, demog2, health, cash, social), collapse = '+')
mod3 = paste0(mod1, "+", mod3)
```

Table 4 Floor (household)

```
# Control group mean & SD for floor list
floor_reg_tbl = data.frame(Variable = character(),
                           Mean_Ctr = numeric(),
                           SD_Ctr = numeric(),
                           M1_Coef = numeric(),
                           M1_RobustSE = numeric(),
                           M1_fraction = numeric(),
                           M2_Coef = numeric(),
                           M2_RobustSE = numeric(),
                           M2_fraction = numeric(),
                           M3_Coef = numeric(),
                           M3_RobustSE = numeric(),
                           M3_fraction = numeric()
                           )
for (i in 1:length(floor)){
  floor_reg_tbl[i,1] = floor[i]
  floor_reg_tbl[i,2] = round(mean(household_ctr[[floor[i]]]), 3)
  floor_reg_tbl[i,3] = round(sd(household_ctr[[floor[i]]]), 3)
}

# Model 1
# Estimated Coefficient for floor
for (i in 1:length(floor)){
```

```
  reg = paste0(floor[i], mod1)
  floor_reg_tbl[i,4] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  floor_reg_tbl[i,5] = round(summary(Pten.gee)$coef[2,4], 3)
}

floor_reg_tbl$M1_fraction = round((floor_reg_tbl$M1_Coef*100)/floor_reg_tbl$Mean_Ctr, 3)

# Model 2
# Estimated Coefficient for floor
for (i in 1:length(floor)){
  reg = paste0(floor[i], mod2)
  floor_reg_tbl[i,7] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  floor_reg_tbl[i,8] = round(summary(Pten.gee)$coef[2,4], 3)
}

floor_reg_tbl$M2_fraction = round((floor_reg_tbl$M2_Coef*100)/floor_reg_tbl$Mean_Ctr, 3)

# Model 3
# Estimated Coefficient for floor
for (i in 1:length(floor)){
  reg = paste0(floor[i], mod3)
  floor_reg_tbl[i,10] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  floor_reg_tbl[i,11] = round(summary(Pten.gee)$coef[2,4], 3)
}

floor_reg_tbl$M3_fraction = round((floor_reg_tbl$M3_Coef*100)/floor_reg_tbl$Mean_Ctr, 3)
floor_reg_tbl
```

Table 6 Satisfaction (household)

```
# Control group mean & SD for satisfaction list
Satis_reg_tbl = data.frame(Variable = character(),
                           Mean_Ctr = numeric(),
```

```
                              SD_Ctr = numeric(),
                              M1_Coef = numeric(),
                              M1_RobustSE = numeric(),
                              M1_fraction = numeric(),
                              M2_Coef = numeric(),
                              M2_RobustSE = numeric(),
                              M2_fraction = numeric(),
                              M3_Coef = numeric(),
                              M3_RobustSE = numeric(),
                              M3_fraction = numeric()
                              )
for (i in 1:length(satis)){
  Satis_reg_tbl[i,1] = satis[i]
  Satis_reg_tbl[i,2] = round(mean(household_ctr[[satis[i]]], na.rm=TRUE), 3)
  Satis_reg_tbl[i,3] = round(sd(household_ctr[[satis[i]]], na.rm=TRUE), 3)
}

# Model 1
# Estimated Coefficient for satis
for (i in 1:length(satis)){
  reg = paste0(satis[i], mod1)
  Satis_reg_tbl[i,4] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  Satis_reg_tbl[i,5] = round(summary(Pten.gee)$coef[2,4], 3)
}

Satis_reg_tbl$M1_fraction = round((Satis_reg_tbl$M1_Coef*100)/Satis_reg_tbl$Mean_Ctr, 3)

# Model 2
# Estimated Coefficient for floor
for (i in 1:length(satis)){
  reg = paste0(satis[i], mod2)
  Satis_reg_tbl[i,7] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  Satis_reg_tbl[i,8] = round(summary(Pten.gee)$coef[2,4], 3)
}

Satis_reg_tbl$M2_fraction = round((Satis_reg_tbl$M2_Coef*100)/Satis_reg_tbl$Mean_Ctr, 3)

# Model 3
# Estimated Coefficient for floor
for (i in 1:length(satis)){
  reg = paste0(satis[i], mod3)
```

```
  Satis_reg_tbl[i,10] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  Satis_reg_tbl[i,11] = round(summary(Pten.gee)$coef[2,4], 3)
}

Satis_reg_tbl$M3_fraction = round((Satis_reg_tbl$M3_Coef*100)/Satis_reg_tbl$Mean_Ctr, 3)
Satis_reg_tbl
```

Table 7 Robustness Checks (household dataset)

```
# Control group mean & SD for Robustness Checks list
Rob_reg_tbl = data.frame(Variable = character(),
                         Mean_Ctr = numeric(),
                         SD_Ctr = numeric(),
                         M1_Coef = numeric(),
                         M1_RobustSE = numeric(),
                         M1_fraction = numeric(),
                         M2_Coef = numeric(),
                         M2_RobustSE = numeric(),
                         M2_fraction = numeric(),
                         M3_Coef = numeric(),
                         M3_RobustSE = numeric(),
                         M3_fraction = numeric()
                         )
for (i in 1:length(robust)){
  Rob_reg_tbl[i,1] = robust[i]
  Rob_reg_tbl[i,2] = round(mean(household_ctr[[robust[i]]], na.rm=TRUE), 3)
  Rob_reg_tbl[i,3] = round(sd(household_ctr[[robust[i]]], na.rm=TRUE), 3)
}

# Model 1
# Estimated Coefficient for robust
for (i in 1:length(robust)){
  reg = paste0(robust[i], mod1)
  Rob_reg_tbl[i,4] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  Rob_reg_tbl[i,5] = round(summary(Pten.gee)$coef[2,4], 3)
}

Rob_reg_tbl$M1_fraction = round((Rob_reg_tbl$M1_Coef*100)/Rob_reg_tbl$Mean_Ctr, 3)
```

```
# Model 2
# Estimated Coefficient for floor
for (i in 1:length(robust)){
  reg = paste0(robust[i], mod2)
  Rob_reg_tbl[i,7] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  Rob_reg_tbl[i,8] = round(summary(Pten.gee)$coef[2,4], 3)
}

Rob_reg_tbl$M2_fraction = round((Rob_reg_tbl$M2_Coef*100)/Rob_reg_tbl$Mean_Ctr, 3)

# Model 3
# Estimated Coefficient for floor
for (i in 1:length(robust)){
  reg = paste0(robust[i], mod3)
  Rob_reg_tbl[i,10] = round(summary(lm(reg, data = household))$coefficients[2, 1], 3)
  # robust se
  Pten.gee = gee(reg,
    id = idcluster,
    family = gaussian,
    corstr = "independence",
    data = household)
  Rob_reg_tbl[i,11] = round(summary(Pten.gee)$coef[2,4], 3)
}

Rob_reg_tbl$M3_fraction = round((Rob_reg_tbl$M3_Coef*100)/Rob_reg_tbl$Mean_Ctr, 3)
Rob_reg_tbl
```

Test for outlierTest

```
library(car)
mod_all = c(mod1, mod2, mod3)
for(i in mod_all){
  print(outlierTest(lm(paste0("S_shcementfloor", i), data=household)))
}
```

Test for Leverage Score

```
# Model 1
# Estimated Coefficient for floor
for (i in 1:length(floor)){
  print(paste(floor[i], "mod1"))
  reg = paste0(floor[i], mod1)
  model = lm(reg, data = household)
```

```r
  hats <- as.data.frame(hatvalues(model))
  hats[order(-hats['hatvalues(model)']), ]
  print(which(hatvalues(model)>0.05))
}

# Model 2
# Estimated Coefficient for floor
for (i in 1:length(floor)){
  print(paste(floor[i], "mod2"))
  reg = paste0(floor[i], mod2)
  model = lm(reg, data = household)
  hats <- as.data.frame(hatvalues(model))
  hats[order(-hats['hatvalues(model)']), ]
  print(which(hatvalues(model)>0.05))
}

# Model 3
# Estimated Coefficient for floor
for (i in 1:length(floor)){
  print(paste(floor[i], "mod3"))
  reg = paste0(floor[i], mod3)
  model = lm(reg, data = household)
  hats <- as.data.frame(hatvalues(model))
  hats[order(-hats['hatvalues(model)']), ]
  print(which(hatvalues(model)>0.05))
}

# Satisfactions
for (i in 1:length(satis)){
  print(paste(satis[i], "mod1"))
  reg = paste0(satis[i], mod1)
  model = lm(reg, data = household)
  hats <- as.data.frame(hatvalues(model))
  hats[order(-hats['hatvalues(model)']), ]
  print(which(hatvalues(model)>0.05))
}
for (i in 1:length(satis)){
  print(paste(satis[i], "mod2"))
  reg = paste0(satis[i], mod2)
  model = lm(reg, data = household)
  hats <- as.data.frame(hatvalues(model))
  hats[order(-hats['hatvalues(model)']), ]
  print(which(hatvalues(model)>0.05))
}
for (i in 1:length(satis)){
  print(paste(satis[i], "mod3"))
  reg = paste0(satis[i], mod3)
  model = lm(reg, data = household)
  hats <- as.data.frame(hatvalues(model))
  hats[order(-hats['hatvalues(model)']), ]
  print(which(hatvalues(model)>0.05))
}
```

Leave One Out CV

```
# LOOCV Error
library(caret)
#specify the cross-validation method
ctrl <- trainControl(method = "LOOCV")
```

FLOOR

```
# floor mod1
for(i in 1:length(floor)){
  modd = as.formula(paste0(floor[i], mod1))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = household, method = "lm", trControl = ctrl,na.action = 'na.omit')
  print(paste(floor[i], "mod1"))
  print(model)
}

# floor mod2
for(i in 1:length(floor)){
  modd = as.formula(paste0(floor[i], mod2))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = household, method = "lm", trControl = ctrl,na.action = 'na.omit')
  print(paste(floor[i], "mod2"))
  print(model)
}

# floor mod3
for(i in 1:length(floor)){
  modd = as.formula(paste0(floor[i], mod3))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = household, method = "lm", trControl = ctrl, na.action = 'na.omit')
  print(paste(floor[i], "mod3"))
  print(model)
}
```

SATIS

```
# satis mod1
for(i in 1:length(satis)){
  modd = as.formula(paste0(satis[i], mod1))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = household, method = "lm", trControl = ctrl, na.action = 'na.omit')
  print(paste(satis[i], "mod1"))
  print(model)
}

# satis mod2
for(i in 1:length(satis)){
```

```
  modd = as.formula(paste0(satis[i], mod2))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = household, method = "lm", trControl = ctrl, na.action = 'na.omit')
  print(paste(satis[i], "mod2"))
  print(model)
}

# satis mod3
for(i in 1:length(satis)){
  modd = as.formula(paste0(satis[i], mod3))
  #fit a regression model and use LOOCV to evaluate performance
  model <- train(modd, data = household, method = "lm", trControl = ctrl, na.action = 'na.omit')
  print(paste(satis[i], "mod3"))
  print(model)
}
```