Barrera-Osorio et al 2011

Dimitrios & Eva

2023-03-08

Contents

1 Motivation

1.1 Why is this research question relevant?

Education in Colombia and other middle-income countries face challenges such as high dropout rates among low-income students, and the reasons behind them, such as the high cost of education. Conditional cash transfers (CCTs) are an evidence-based intervention to increase participation in education. However, the authors highlight that there is little variability in the structure of programmes. The paper investigates if changes in the timing of payments affect the outcomes of interest: Attendance and re-enrollment. Optimising the structure of CCTs may contribute to improved education outcomes and reduce disparities in access to education.

1.2 What are the main hypotheses?

- The savings model will improve outcomes compared to the basic programme by relaxing possible savings constraints.
- The tertiary model will improve rates of graduation and tertiary enrollment compared to the basic programme by providing direct incentives for continuation of education.

2 Data sources

We investigated performing the replication with the data provided as part of the lecture. However, we soon discovered that the file did not contain all required variables, nor was there any meta data or other information about the variables in the dataset. A brief search revealed that the data and STATA scripts used to obtain the authors' results are freely available here.

For this project, we reference the following files: * Data file: Public_Data_AEJApp_2010-0132.dta * STATA script for Table 3: Table_03_Attendance.do * Meta data: AEJApp_2010-0132_Data_ReadMe.pdf

2.1 Where does the data come from (country, time period, source)?

Add

2.2 What are the key variables and ow are these measured?

Add

3 Method

3.1 Research design

The research paper describes three interventions designed to improve attendance and educational outcomes for students in Colombia.

The first intervention ("basic") is similar to the PROGRESA/OPORTUNIDADES program, a conditional cash transfer program in Mexico that operated from 1997 to 2012. It pays participants 30,000 pesos per month (approximately US$15) if the child attends at least 80% of the days in that month. Payments are made bi-monthly through a dedicated debit card, and students will be removed from the program if they fail to meet attendance targets or are expelled from school.

The second intervention, called the savings treatment, pays two-thirds of the monthly amount (20,000 pesos or US$10) to students' families on a bi-monthly basis, while the remaining one-third is held in a bank account. The accumulated funds are then made available to students' families during the period in which students prepare to enroll for the next school year, with 100,000 pesos (US$50) available to them in December if they reach the attendance target every month.

The third intervention, called the tertiary treatment, incentivizes students to graduate and matriculate to a higher education institution. The monthly transfer for good attendance is reduced from 30,000 pesos per month to 20,000 pesos, but upon graduating, the student earns the right to receive a transfer of 600,000 pesos (US$300) if they enroll in a tertiary institution, and after a year if they fail to enroll upon graduation.

All payments are based on reports provided to the Secretary of Education by the students' principals. Students will be removed from the program if they fail to meet attendance targets, fail to matriculate to the next grade twice, or are expelled from school.

The eligibility criteria for the basic and savings experiments were as follows:

- Children had to have finished grade 5 and be enrolled in grades 6 - 11.
- The children's families had to be classified into the bottom two categories on Colombia's poverty index (SISBEN).
- Only households living in San Cristobal prior to 2004 were eligible to participate.

The eligibility criteria for the tertiary education experiment were as follows:

- Children had to have finished grade 8 and be enrolled in grade 9 - 11.
- The children's families had to be classified into the bottom two categories on Colombia's poverty index, the SISBEN.
- Only households living in Suba prior to 2004 were eligible to participate.

The paper investigates differences in enrollment and graduation / progression to tertiary education for the three treatment groups compared to untreated controls. Randomization to treatment vs control group was stratified by location, school public vs private, gender and grade.

3.2 Data preparation

some text here

ADD: Steps we've taken to use the data. The dataset for our analysis is called "filtered_barrera" (put this to R-code).

3.3 Analysis

3.3.1 What are the assumptions of the method?

The authors initially use simple linear regression to compare treatment groups. They model the relationship between a dependent variable (outcome; attendance) and two independent variables (whether participant is allocated to treatment "basic", and whether participant is allocated to treatment "savings".)

The assumptions about the data underlying linear regression are:

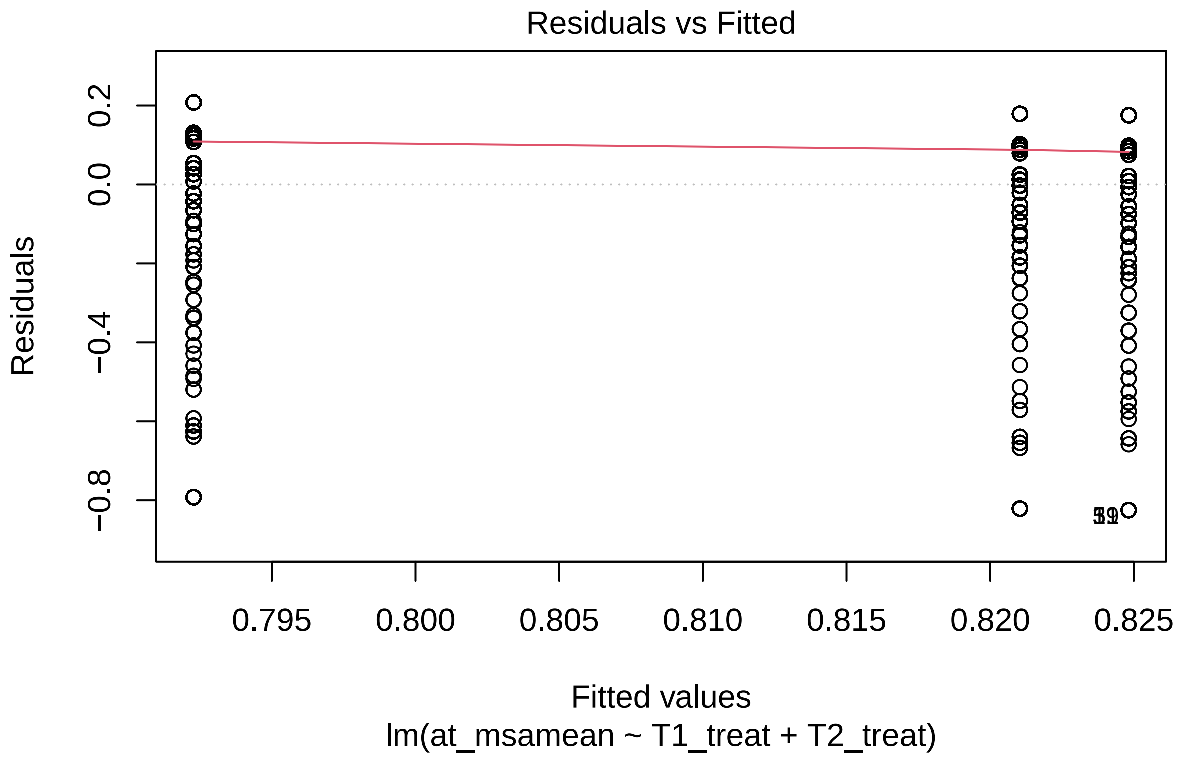
1. Linearity: There should be a linear relationship between the independent and dependent variables.
2. Independence: The observations used in the regression analysis should be independent of each other. In other words, the value of one observation should not be influenced by the value of another observation.

3. Homoscedasticity: The variance of the dependent variable should be constant across all values of the independent variable(s).
4. Normality: The dependent variable should be normally distributed at each level of the independent variable(s).
5. No multicollinearity: If there are multiple independent variables in the regression model, there should be no high correlation between these independent variables.

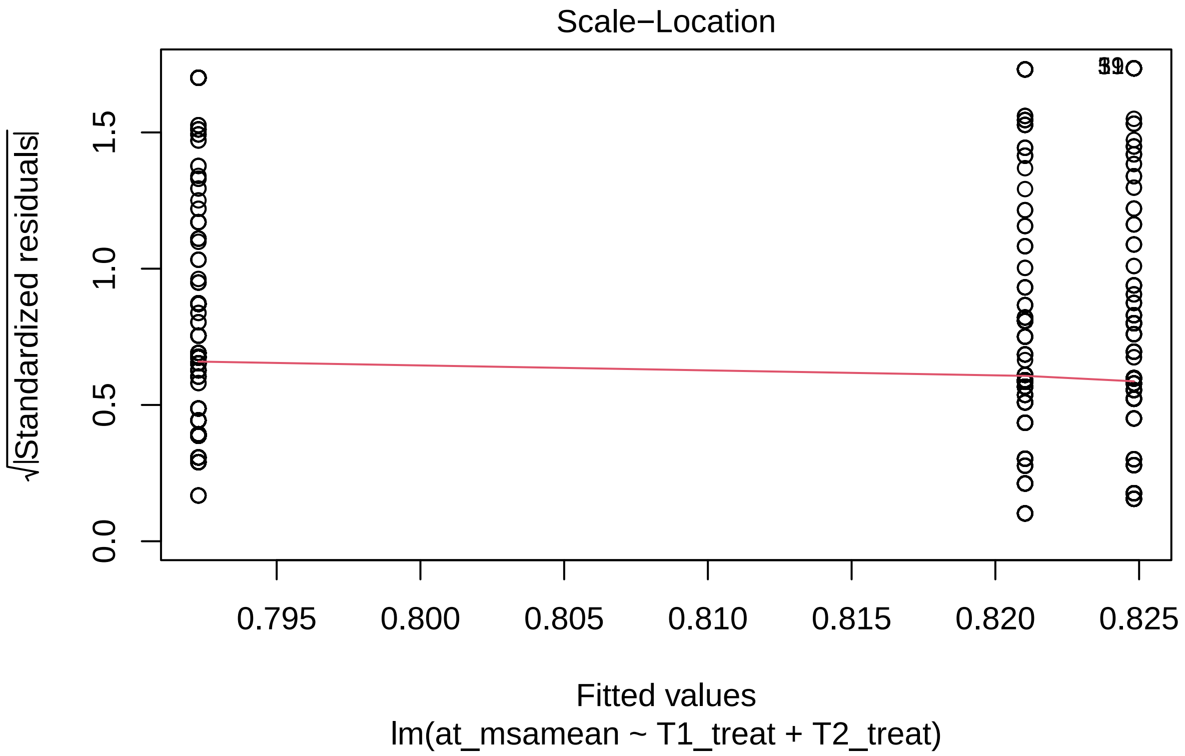
If these assumptions are not met, this can lead to unreliable estimators (regression coefficients) and / or biased standard errors, i.e. standard errors that are systematically smaller or larger than the "true" standard error. This means that the relationship between dependent and independent variables is not estimated correctly by the model.

3.3.2 Are these assumptions plausible in this example?

We test the assumptions of the most simple model using the procedure detailed here:

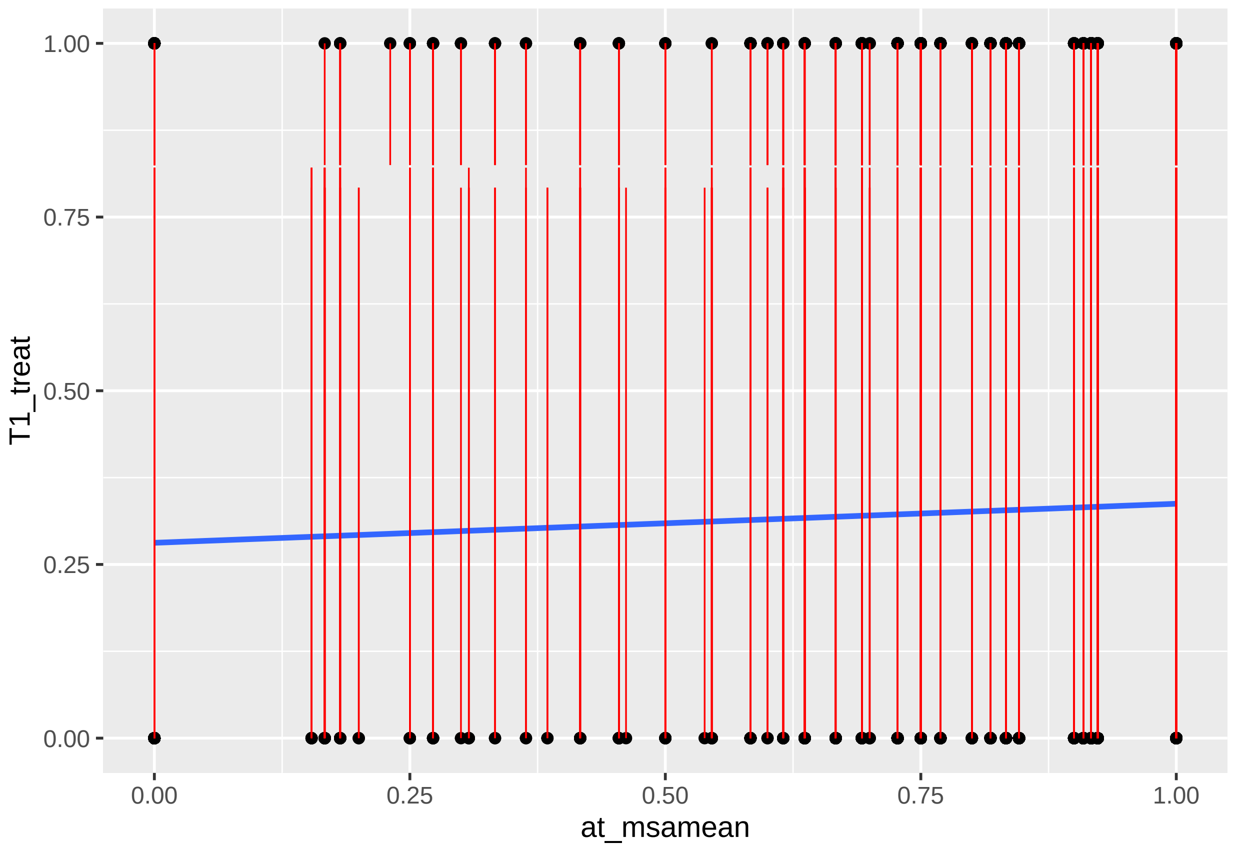


```
##  lag Autocorrelation D-W Statistic p-value
##    1     0.004010807      1.990513   0.698
##  Alternative hypothesis: rho != 0
```



```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



One violation that should be expected based on the data is that of independence. Observations are clustered within schools and also individuals. The authors address this issue by clustering error terms at the school level:

Model 1:

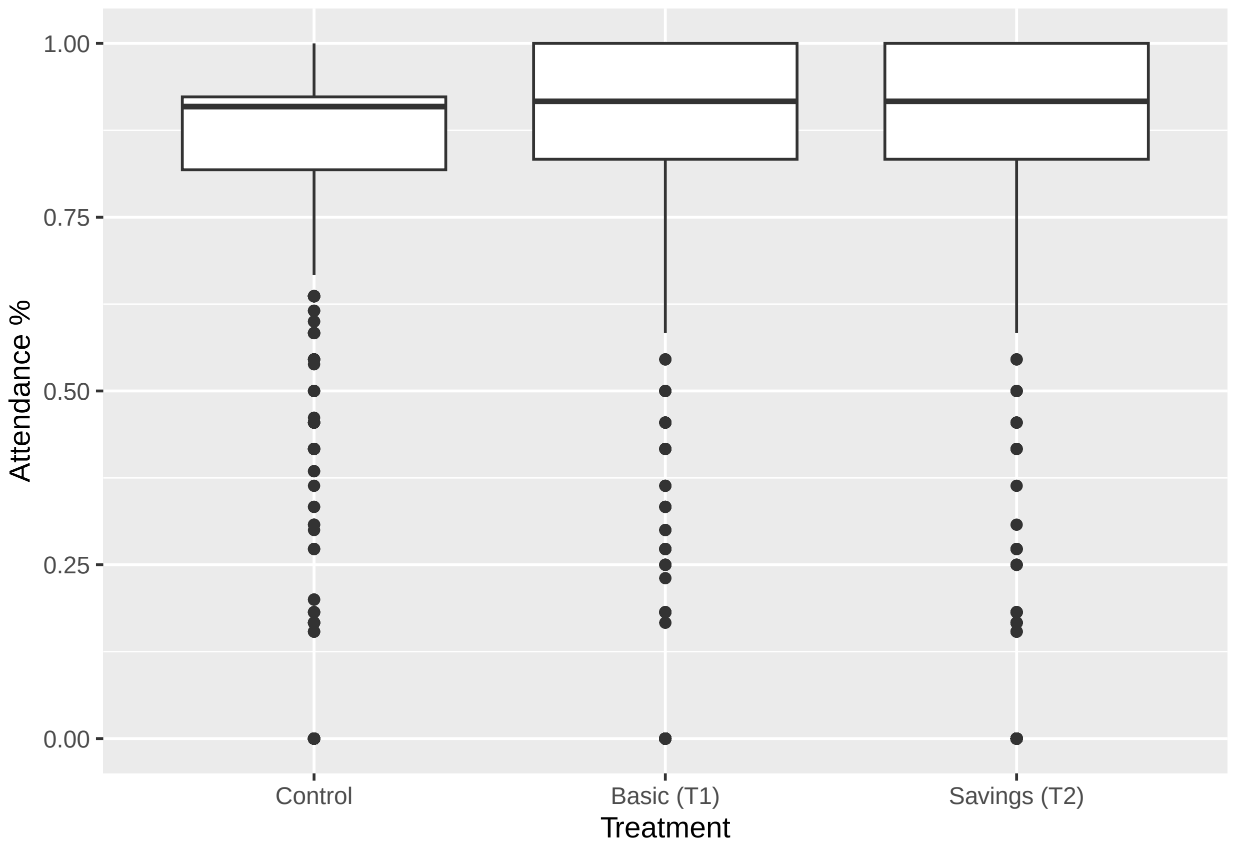
$$y_{ij} = \beta_0 + \beta_B Basic_i + \beta_S Savings_i + \epsilon_{ij}$$

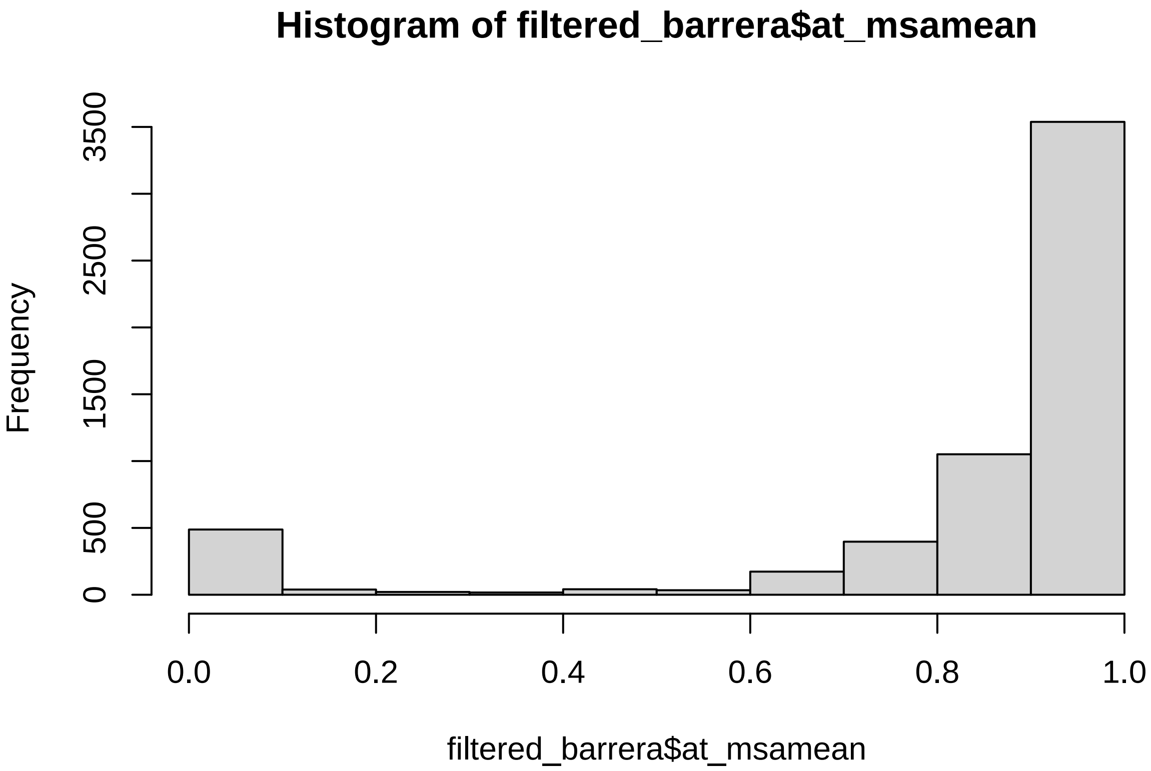
Model 2:

$$y_{ij} = \beta_0 + \beta_B Basic_i + \beta_S Savings_i + \delta X_{ijk} + \theta_j + \epsilon_{ij}$$

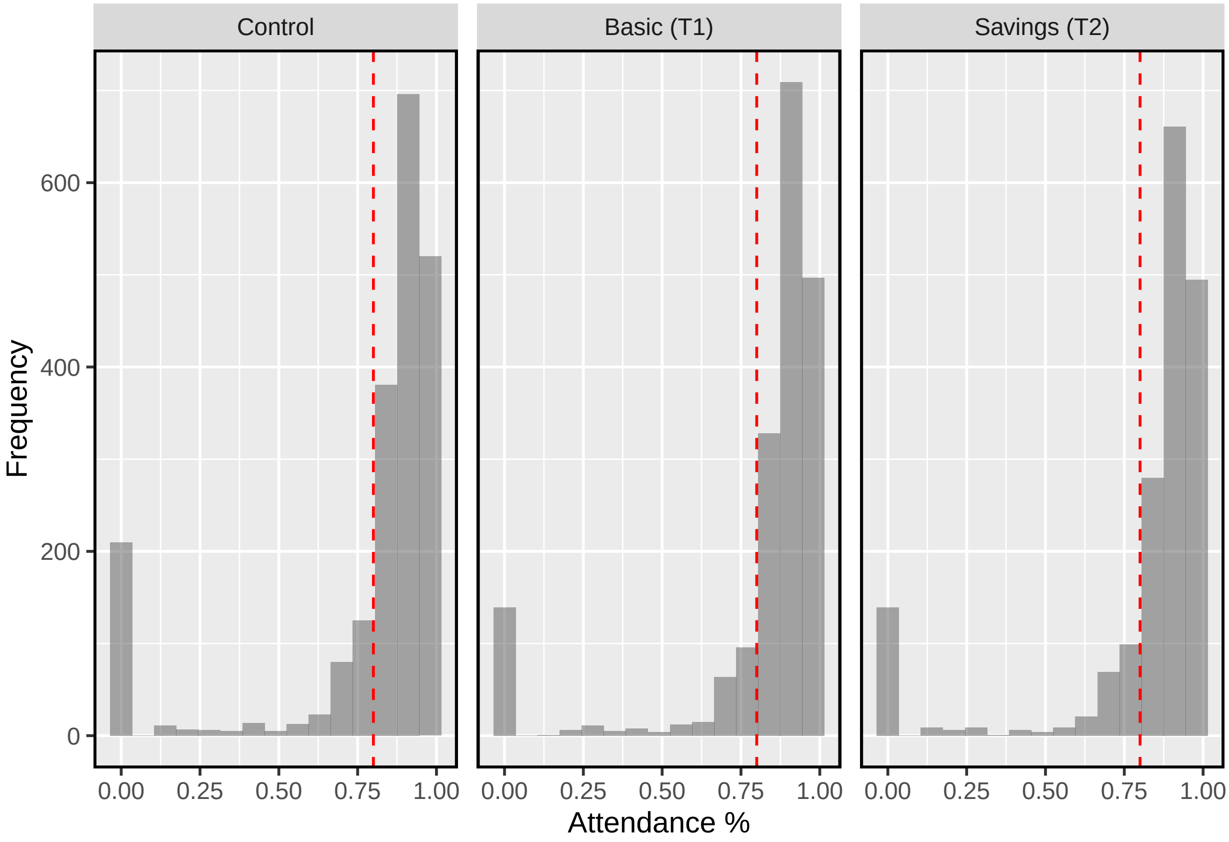
Fixed effect models: Used to control for unobserved factors that are constant over time at any level of analysis. The key is to identify the appropriate level of analysis for the fixed effect and include it in the model to account for the unobserved factors that affect the outcome variable. In this case, the school was chosen as the fixed effect, while standard errors are clustered within the individual.This reflects the multiple levels of clustering off effects (unobserved school characteristics and unobserved characteristics of individuals).

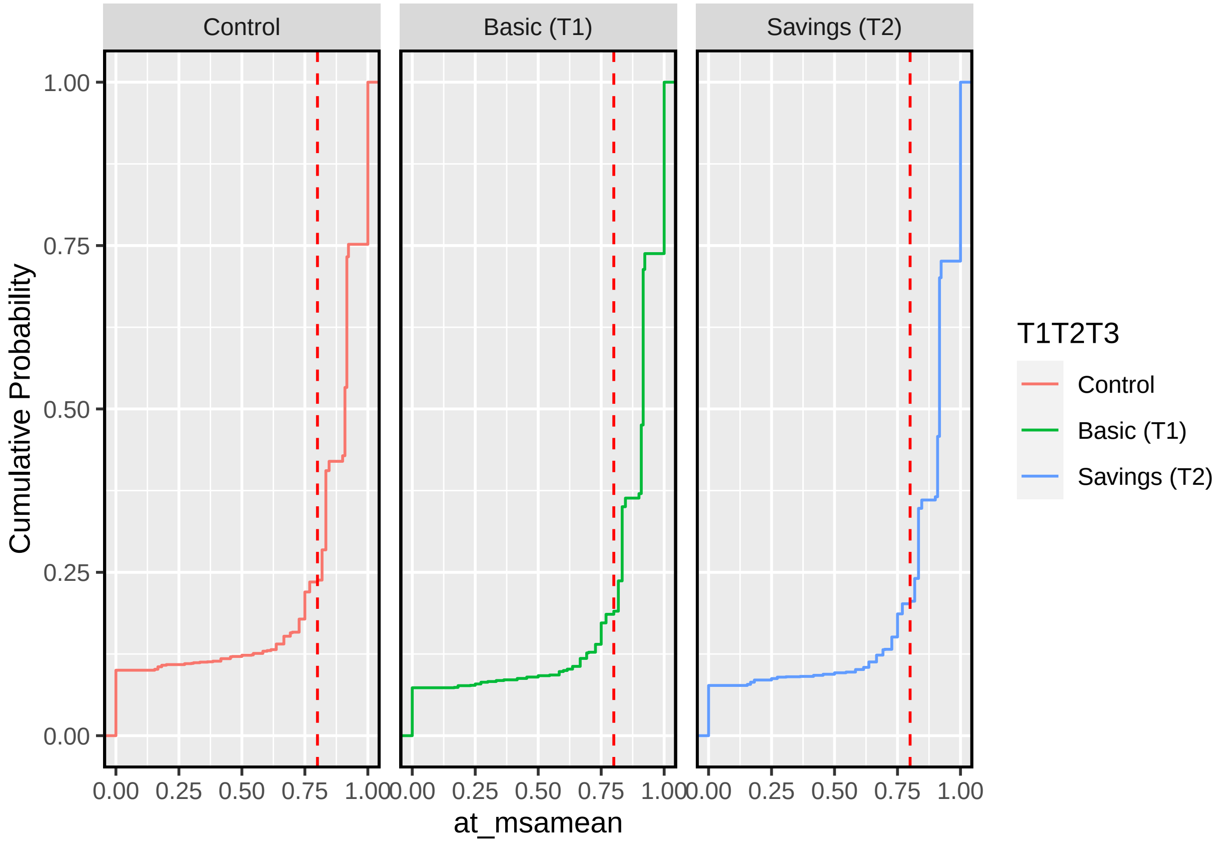
Problem: While there are some plots built into R, a lot of these just don't work easily with clustered data. My workaround: Plotting on the simplest model (model 1) to show that simple linear regression would not have been a good choice. But that seems very obvious.


1.00
0.75
0.50
0.25
0.00
Attendance %
Control
Basic (T1)
Savings (T2)
Treatment



```
## Warning: The ‘size‘ argument of ‘element_rect()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
```


Control
Basic (T1)
Savings (T2)
Frequency
600
400
200
0
0.00
0.25
0.50
0.75
1.00
Attendance %



```
## # A tibble: 3 x 2
##   T1T2T3       prop_cutoff
##   <fct>              <dbl>
## 1 Control            0.765
## 2 Basic (T1)         0.814
## 3 Savings (T2)       0.798
```

4 Descriptive statistics

- Describe sample

```
# Table 1




# Using n = 5,799 to get the sample actually used in our model (for columns 1-3). Variables selected ba
# Note: Some factor variables are presented in the paper as scale vars.

# House posessions - f_teneviv
# utilities - s_utilities
# durable goods - s_durables
# physical infrastructure - s_infraest_hh
# age - s_age_sorteo
# gender - s_sexo
# years of education - s_yrs
# single head - s_single
# Age of head - s_edadhead
# years of ed head - s_yrshead
# people in household - s_tpersona
# Member under 18 - s_num18
# estrato - f_estrato
# SISBEN - s_puntaje
# household income - s_ingtotal


table1 <- table1(~ factor(f_teneviv) + s_utilities + s_durables + s_infraest_hh + s_age_sorteo + factor

table1
```

	Control	Basic (T1)	Savings (T2)	Overall
	(N=2096)	(N=1895)	(N=1808)	(N=5799)
factor(f_teneviv)				
1	1160 (55.3%)	1013 (53.5%)	992 (54.9%)	3165 (54.6%)
2	150 (7.2%)	133 (7.0%)	141 (7.8%)	424 (7.3%)
3	570 (27.2%)	524 (27.7%)	503 (27.8%)	1597 (27.5%)
4	216 (10.3%)	225 (11.9%)	172 (9.5%)	613 (10.6%)
s_utilities				
Mean (SD)	4.64 (1.42)	4.62 (1.40)	4.69 (1.39)	4.65 (1.40)
Median [Min, Max]	5.00 [1.00, 6.00]	5.00 [1.00, 6.00]	5.00 [1.00, 6.00]	5.00 [1.00, 6.00]
s_durables				
Mean (SD)	1.35 (0.883)	1.32 (0.881)	1.39 (0.871)	1.35 (0.879)
Median [Min, Max]	1.00 [0, 4.00]	1.00 [0, 4.00]	1.00 [0, 4.00]	1.00 [0, 4.00]
s_infraest_hh				
Mean (SD)	11.6 (1.75)	11.5 (1.82)	11.7 (1.63)	11.6 (1.74)
Median [Min, Max]	12.0 [3.00, 19.0]	12.0 [3.00, 18.0]	12.0 [3.00, 17.0]	12.0 [3.00, 19.0]
s_age_sorteo				
Mean (SD)	14.1 (5.42)	14.2 (5.56)	13.9 (5.04)	14.1 (5.35)
Median [Min, Max]	13.0 [4.00, 72.0]	13.0 [1.00, 76.0]	13.0 [3.00, 78.0]	13.0 [1.00, 78.0]
factor(f_sexo)				
Female	1055 (50.3%)	931 (49.1%)	916 (50.7%)	2902 (50.0%)
Male	1041 (49.7%)	964 (50.9%)	892 (49.3%)	2897 (50.0%)
Years of Education				
Mean (SD)	5.34 (1.72)	5.27 (1.70)	5.29 (1.69)	5.30 (1.70)
Median [Min, Max]	5.00 [0, 14.0]	5.00 [0, 16.0]	5.00 [0, 12.0]	5.00 [0, 16.0]
factor(f_single)				
No	1492 (71.2%)	1334 (70.4%)	1270 (70.2%)	4096 (70.6%)
Yes	604 (28.8%)	561 (29.6%)	538 (29.8%)	1703 (29.4%)
Age of Jefe del Hogar				
Mean (SD)	45.6 (10.3)	45.5 (9.74)	45.8 (9.80)	45.6 (9.97)
Median [Min, Max]	43.0 [19.0, 91.0]	43.0 [23.0, 98.0]	44.0 [24.0, 84.0]	43.0 [19.0, 98.0]
Years of Education Jefe				
Mean (SD)	5.59 (2.92)	5.56 (2.80)	5.49 (2.89)	5.55 (2.87)
Median [Min, Max]	5.00 [0, 22.0]	5.00 [0, 15.0]	5.00 [0, 16.0]	5.00 [0, 22.0]
Number of people in the household				
Mean (SD)	5.40 (1.94)	5.44 (1.93)	5.41 (1.93)	5.42 (1.93)
Median [Min, Max]	5.00 [2.00, 19.0]	5.00 [2.00, 19.0]	5.00 [2.00, 19.0]	5.00 [2.00, 19.0]
Number of kids 18 and under				
Mean (SD)	2.63 (1.32)	2.71 (1.35)	2.66 (1.33)	2.67 (1.33)
Median [Min, Max]	2.00 [0, 11.0]	3.00 [0, 12.0]	2.00 [0, 12.0]	2.00 [0, 12.0]
factor(f_estrato)				
0	440 (21.0%)	408 (21.5%)	379 (21.0%)	1227 (21.2%)
1	292 (13.9%)	256 (13.5%)	267 (14.8%)	815 (14.1%)
2	1364 (65.1%)	1231 (65.0%)	1162 (64.3%)	3757 (64.8%)
SISBEN score				
Mean (SD)	11.7 (4.64)	11.5 (4.51)	11.5 (4.52)	11.6 (4.56)
Median [Min, Max]	12.4 [1.92, 21.9]	12.4 [2.28, 21.8]	12.3 [1.82, 22.0]	12.3 [1.82, 22.0]
Household Income				
Mean (SD)	367 (239)	358 (240)	368 (226)	364 (235)
Median [Min, Max]	332 [0, 3320]	330 [0, 4000]	332 [0, 1730]	332 [0, 4000]

5 Results

- Are these results plausible?
- How robust are the results to changing the sample?

```
## The variables 'f_estrato3', 'f_grade10' and two others have been removed because of collinearity (se
## The variables 'f_estrato3', 'f_grade10' and two others have been removed because of collinearity (se
```

	(1)	(2)	(3)
T1_treat	0.033	0.032	0.032
	(0.007)	(0.008)	(0.007)
T2_treat	0.029	0.027	0.027
	(0.008)	(0.008)	(0.007)
Num.Obs.	5799	5799	5799
R2	0.003	0.037	0.089
R2 Adj.	0.003	0.032	0.080
R2 Within			0.037
R2 Within Adj.			0.032
AIC	1447.1	1302.3	1030.3
BIC	1467.1	1502.3	1416.9
RMSE	0.27	0.27	0.26
Std.Errors	by: school_code	by: school_code	by: school_code
FE: school_code			X

6 Graph

ggplot(data=filtered_barrera, aes(x=at_baseline, y=at_msamean, color=factor(T1T2T3))) + geom_point()
+ geom_smooth(method="lm", se=FALSE)

ggplot(data=filtered_barrera, aes(x=at_baseline, y=at_msamean, color=factor(T1T2T3))) + geom_smooth(method="lm"
se=FALSE)+ xlim(0.65, NA) + ylim(0.5, NA)

7 Some exploration

8 Part of the problem with this model is the outcome variable, which has a ceiling effect (can't go above 100%, and many people have high attendance, with target attendance also being high at 80%.)

9 Alternative way of approaching this:

10 GLM for skewed data (e.g. log link and gamma function –> would need to fit this more carefully,)

11 Binary variable: Whether or not student achieved 80% attendance

12 Calculating new column: is at or above cut-off?

```
filtered_barrera <- filtered_barrera %>% mutate(above_cutoff = ifelse(at_msamean >= cutoff, 1, 0))
```

13 Running above model but with this as outcome –> note that the residual plots don't pick up the clustered standard errors.

```
library(clusterSEs) # to get augment method for lm.cluster library(sandwich) library(broom) # augment for glm
```

```
mod_bi <- glm(data = filtered_barrera, above_cutoff ~ T1_treat + T2_treat + f_teneviv + s_utilities + s_durables + s_infraest_hh + s_age_sorteo + s_age_sorteo2 + s_years_back + s_sexo + f_estcivil + s_single + s_edadhead + s_yrshead + s_tpersona + s_num18 + f_estrato + s_puntaje + s_ingtotal + f_grade + suba + s_over_age + factor(school_code), family = binomial()) summary(mod_bi) vcov <- vcovCL(mod_bi, cluster = filtered_barrera$school_code) coeftest(mod_bi, vcov = vcov)
```

```
fitted_bi <- augment(mod_bi, data = filtered_barrera, se_fit = TRUE)
```

```
plot_bi <- ggplot(fitted_bi, aes(x = .resid)) + geom_histogram(binwidth = 0.01, color = “white”, boundary = 50000) plot_bi
```

14 To compare, running linear model as glm:

```
mod_gau <- glm(data = filtered_barrera, at_msamean ~ T1_treat + T2_treat + f_teneviv + s_utilities + s_durables + s_infraest_hh + s_age_sorteo + s_age_sorteo2 + s_years_back + s_sexo + f_estcivil + s_single + s_edadhead + s_yrshead + s_tpersona + s_num18 + f_estrato + s_puntaje + s_ingtotal + f_grade + suba + s_over_age + factor(school_code), family = gaussian()) summary(mod_gau) vcov <- vcovCL(mod_gau, cluster = filtered_barrera$school_code) coeftest(mod_gau, vcov = vcov)
```

```
#fitted_gau <- augment(mod_gau, data = filtered_barrera) fitted_gau <- augment(mod_gau, data = filtered_barrera, se_fit = TRUE)
```

plot_gau <- ggplot(fitted_gau, aes(x = .resid)) + geom_histogram(binwidth = 0.01, color = “white”, boundary = 50000) plot_gau

library(gridExtra) # to show two plots next to each other - I know there is another way! # arrange plots side by side grid.arrange(plot_bi, plot_gau, nrow = 2)

15 GLM with binary outcome variable performs much better on AIC. But I’m not sure the residual plots help us very much as it’s not clustered SE!

16 Conclusion

- Replication of a research paper: How do results compare to results of research paper?

Models all bad, R-sq low.

Compare coefficients, standard errors Why might they be different? * Software: Possible to get STATA standard errors in R

Source for clustered standard errors in R: https://evalf21.classes.andrewheiss.com/example/standard-errors/