Econometrics I - Empirical Work

Ricardo Semião e Castro

04/2024

Introduction

Before going to the questions, I'll present some topics on the empirical strategy that are relevant to the whole exercise.

The empirical papers in the literature often do some kind of filtering of the data. There are a lot of different types of households and families, there is a important trade-off in cleaning "bad" cases, such as families with absent parents, step children, old "children", etc. On one hand, filtering removes noise and biases, helping the interpretation of the results; but, on the other (i) we lose observations, which yield less precise inference, (ii) reduces the population being studied and the extrapolation potential, and (iii), if poorly made, might induce a selection bias and/or be seen as arbitrary.

In the present case, the selected census, of Puerto Rico 2000, had $13,372$ households, with only $6,512$ with children, such that item (i) is really relevant, some "bad" cases were too costly to remove. Below I present the filtering strategy used in this exercise:

- Households without children were removed.
  - Defining children as descendants under 18 years old (the "medium" legal age in Puerto Rico), there are 25.48% of original households with children. This is a very low sample size, so it was considered children as descendants under 21 years old (the full legal age in Puerto Rico), but that aren't yet married nor working, such that they're probably still accounted by the parents in the development decision. We're left with 28.21% of original cases, $3,970$ households. The percentages below refer to this baseline.
- Households with more than one family were removed (1.66% of cases), as it is hard to argue how is the development decision made, at the family or at the household level;
- Households classified as "group quarters" were removed (0.1% of cases), for the same reason as above;
- Households with absent mothers were removed (3.33% of cases).
  - Absent fathers were a lot more common (33.4% of the cases), and had to be kept. The literature seems to give more focus on the mother's role in the child's development, such that keeping households with absent fathers is less problematic. I

also discuss including a control for this;

  - Households with step parenting were kept (7.54% of the cases) as it was assumed that different parenting status didn't changed too much the development decision. I also discuss including a control for this;

- Households where there were descendants of the head of the family with more than 21 years old were removed, even tough they had a large presence (12.53% of original cases). In these households, the old descendants enter completely differently in the development decision, such that a lot of noise comes from these cases.
- Households where there were simultaneously children and grandchildren were removed (9.58% of the cases). It is not clear if a family counts the grandchildren as children in the development decision.

One important filtering to be made was to remove families with children already out of the household. Similar to the old descendants, such cases highly impact the development decision and create a lot of noise. But, the census data for Puerto Rico doesn't have a variable for number of children, only for number of children at the household. This is a big limitation of this study. Children out of the household are unaccounted for, which probably induces a underestimation effect.

The low sample size, combined with the aforementioned underestimation effect, might decrease the statistical significance of the findings in this exercise.

Remark: all of the R and Latex code related to this work can be found in my Github.

Question 1

There are a lot of ways that the income can be related with the family size. The most simple one is a linear relation, where richer families have the resources to have more children (positive relation), or where poor families have more children because they need the labor to maintain the households (negative relation).

One would imagine that both these stories might be true, such that the relation is non-linear. There might be a negative effect but with a increasing "return" of income on family size, a polinomial relation, where poor families have lots of children, middle class tend to have very little, and richer families more children too.

But, it could be the case that ever increasing income does not reflect into ever increasing family size. The effect might be binned/discontinuous, such that each bin of income has a different average of children, and even a different effect of income on family size, such that for really rich families, increases in the income don't relate to increases in family size.

In the same sense of the distribution (bins) of income being important, it might be that it is not the absolute value of income that matters, but actually the relative position of the family in the income distribution. For example, it might be the access to certain areas and benefits of society that is more related to the number of children, such that the CEF is related to the quantiles of income.

Question 2

The models considered to explain the number of siblings were:

- Simple linear specification:
  - A linear term of total income;
- Binned specifications:
  - A linear term of total income, plus an intercept for each bin of income;
  - The same as above, but with a different slope for each bin too;
- Linear + non-linear term specification:
  - A linear plus a quadratic term of total income;
  - A linear plus a log term of total income;
- CDF based specifications:
  - An accumulated density of total income;
  - The same as above but also with a quadratic term.
  - The same as above but binned effects.

The linear specification is the most basic one. The binned specifications consider a different mean for each bin, and the second option, also letting the effect (slope) itself vary for each bin. This is in line with trying to saturate the CEF[1].

The cuts were chose to separate classes of yearly income in the puerto rican population: 0-$25,000$\$/year, $25,000$-$50,000$\$/year, $50,000$-$100,000$\$/year, and $100,000$ onwards.

The linear + non-linear term specification tries to capture diminishing or increasing effects, the quadratic version in a polinomial fashion, and the log version in an exponential fashion.

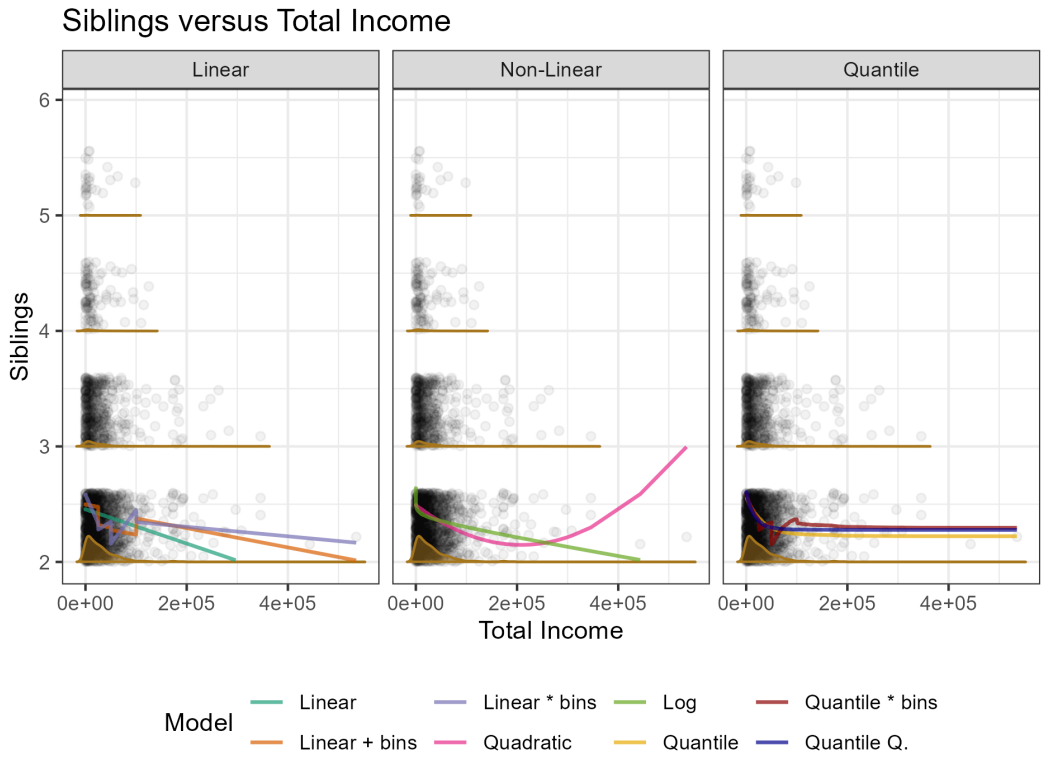
Lastly, the CDF based specifications are a way to model the CEF with the relative position of income being the defining factor for the association with the number of siblings. There might also be a diminishing or binned effect in this scenario too.

The results are plotted below. The regression tables can be seen in the section appendix.

The linear model and quadratic have their obvious limitations, with a very strange fit for large values in the edge of the data. The binned quantile model seems to suffer from "overfitting".

We can see that most of the models have some sense of large number of siblings for low income families, then a initial decrease for middle income families. The linear based and quadratic models present different opposite relations for the high income families, but the log and quantil based approaches all show a stabilization around 1.75 children. That is, it seems that, for high income, increases in income don't relate to different number of children.

[1] But it is actually impossible to do so for continuous variables and finite observations.



Question 3

Note that, even if the CEF does not present the exact format as one of the presented, the OLS estimator would still find the best approximation that the format can give, such that we could have a causal interpretation even without perfect matching of the true CEF.

The truly relevant question is as follows: the CEF having causal interpretation require that the used income measure is exogenous. That is, in relation to a potential outcomes (populational) model, its error term – which holds everything that is relevant to the definition of the family size that isn't income – must be uncorrelated with the income measure.

This is probably false in our case. Any of the controls talked about in question 8 are variables that are relevant to the family size definition, and have a probable relation to income. To quote some stories: families with older children might already hit the wanted mark of a larger family, and don't have the negative effect on income that is to take care of small children; families where the parents have a higher education probably have higher income, and there might be a trend of a given family size amongst those parents.

We can even find evidence for the correlation of income and the other variables. For example, in our sample, the correlation between income and the average age of the parents is 0.45, between income and the average education of the parents is 0.42, between income and a dummy for still together parents is 0.38.

Question 4

Again, we can have different formats to such CEF. The most basic one would be linear, where each increased sibling is related to a lower quality of the rest (negative relation, in line with the literature), or where more siblings create positive spillovers onto the rest of the family (positive relation).

We can also consider non-linear relation, where the relation of siblings and quality is not constant, starts negative but presents "increasing effects", or the other way around. The stories that would justify such relations require a definition of the quality measure. For example, if the quality is "bedrooms per child", it might be that there is the intuitive negative relation at first, but really large families put a increased importance in bigger homes.

The most important topic on this CEF is the fact that Siblings is a discrete variable, such that the CEF might be a different point for each possible value of family size. A OLS regression with a categorical siblings variable would completely staturate the CEF.

Question 5

The relevant question is "how can a setting influence the quality-per-child decision?". Settings where there is a lot of guaranteed quality for children, regardless of the family size, should present CEFs with a less intense relation. These cases can include countries and historical periods with more welfare programs such as public schooling and health. The opposite is true, in settings with less guaranteed welfare, we would expect a more intense relation.

In settings where the necessity of household labor is big, we would expect a more negative relation, as the families choose to have more children even at the cost of quality. So poorer, perhaps less industrialized settings will present this kind of relation. The opposite is true, when the decision of having children is less financial and more about will of the family, the relation should change.

In summary, any fact that changes the development decision can define settings where the relation is different.

Question 6

The measures of quality considered were:

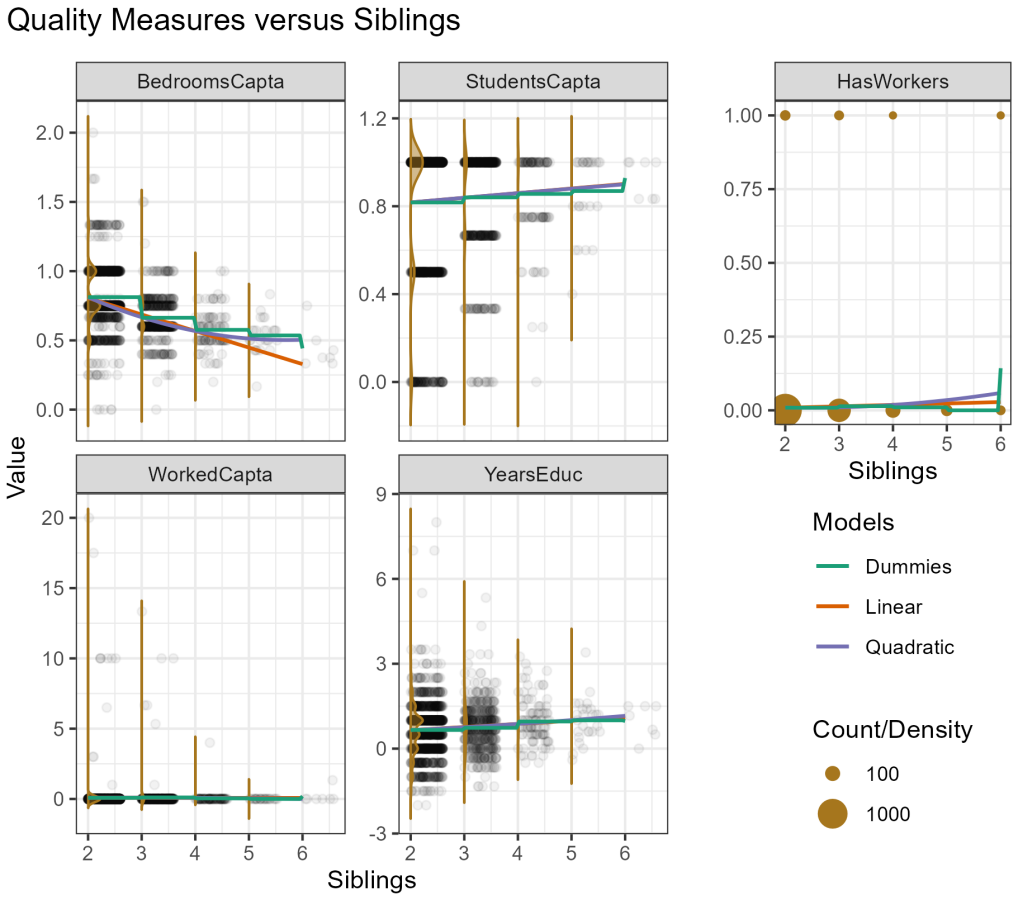
- Number of bedrooms in the house per child: a measure of home quality, parents that want good development for their children will increase the size of the home as the size of the family increases.
- Distance from ideal schooling years: in Puerto Rico, children should start schooling at the age of 5, delayed children will have a negative value, and the opposite for advanced children. See the appendix for more information.
- Number of enrolled students per child: another measure of education quality, households where the parents enrolled all the children in a school, its value will be 1, and

lower otherwise.

- Number of children working per child: a negative measure of quality, if all the children are enrolled in child labor programs, the value will be 1, and lower otherwise.

Three different specifications were considered for the regression on number of siblings: a fully saturated model with dummies for each possible value (1 - 7), a linear polinomial, and a polinomial.

The results can be seen below. The regression tables (for the most relevant measures) can be seen in the section appendix.



First of all, we see that the quadratic polinomial matches really well the fully saturated model.

The rooms per capta presents a negative relation, as expected, but we know that part of the effect is a simple logistic problem of the rigidity of increasing the number of bedrooms in a house or changing houses, such that this is not the most interesting measure. Importantly, it might not react to unexpectedly changes in the family size. It will be kept as a probably upward (in absolute terms) biased effect.

The distance from ideal schooling years presents a very weak relation, slightly positive, the opposite of what was expected. Still, it is the best measure, and lets analise what happens when we control for confounders ahead.

The other measures aren't too interesting. The frequency of child labor is really small, and "students per capta" is a discrete measure less informative than the distance from ideal schooling years.

Not restricting the sample for certain families, such as families where the oldest son is 14 years old, can be a problem of comparability of some variables between households. At the same time, recall the discussion in the introduction about problems with sample restrictions. As the sample size is already small, the full sample was kept. In order the decrease the comparability problem, only measures that did not depend on the family ages were considered. For example, the measure "years of schooling" wouldn't be comparable, and wasn't considered.

This is not ideal, and the same worry was present in the selection of controls ahead, but the sample restriction would be too costly.

Question 7 and 8

Remark: I choose to answer both jointly, for better organization and less repetition.

As it was commented earlier, the requirement to causal interpretaion is no correlation to any ommited relevat variable to the quality of the children.

This again is probably false. The same controls from question 8 are variables are relevant to the developemtn decision too. Consider the control variables quoted below, each had a quick justification, normally on the style of "why does it relate to the number of siblings" and "why is it relevant for the quality measures", such that all are variables that would hurt the causal interpretation of the CEF.

- Children Controls:
  - Frequency of male children: there is a literature on the effect of having children of a same or specific gender on the family size. Also, the gender of the children might be related to the development decision, specially the school attendance;
  - Average children age: "older families" are more probable to have already reached the planned number of children. Also, the development decision might be different for children of different ages;
  - Age of youngest child: younger children might be more present in larger families. Also, parents that have to take care of small children might focus less on the quality of the rest of the household;
  - Presence of disabled children: families that had a disabled child might have a different plan for family size. Also disabilities pose some chalenges in school attendance;
- Family Controls:
  - Non-atomic family size: the number of non parents nor children in the household. Other adults might affect both the family size plan and the development decision;
  - Income (linearly and as quantiles): as was seen, the income of the household is

one of the most important controls. The quantile version was chosen for its good fit and intuition, as presented in question 2;

  - Percentage of welfare income: the percentage of the family income that comes from welfare programs. It helps to better capture the effects of income, specially the different effects for the poorer families;
  - Percentage of pension income: the percentage of the family income that comes from pension. Important for a similar reason as above;

- Parents Controls:
  - Some characteristics of the parents affect the ability, interest, and time that they have available to grow big families and induce a good development for the children, such as average parents age, average parents education, and average parents work hours. The average was used as not always the father is present.
  - It is also important to control for the cultural and social context of the family, such as parents' race and parents' citizenship. Both are the race/citizenship status if both parents present the same, or a value for biracial/bi-citizenship relation.
  - Lastly, we include a control for families where at least one of the parents are step-parents. Step parents might give a different importance to family size and development quality.

We can also find the sample correlations between those and the number of siblings. For example. age of the yougest child has a correlation of -0.135, worked hours of parents has a correlation of -0.152. Again, there is strong reasoning to not interpret the cited CEF as causal.

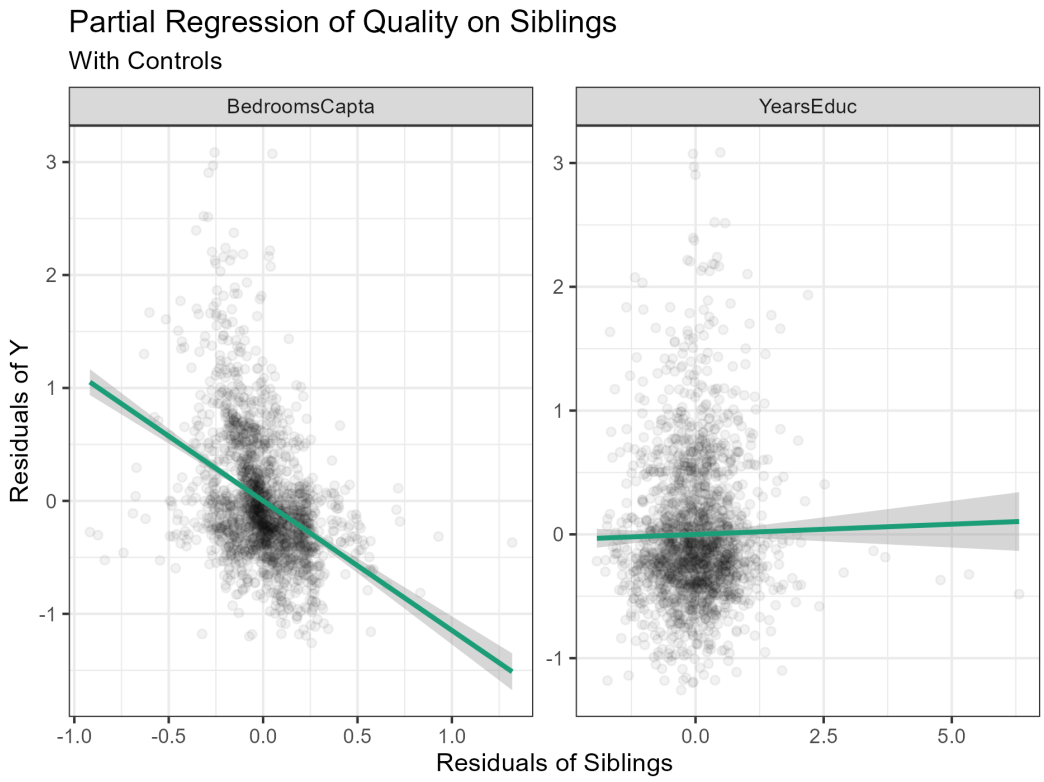
A last relevant variable is the presence of the father in the household, as it was discussed in the introduction. But, including it would be a problem, as it can be that increases in the number of siblings causes the father to become absent, such that it might be a bad control. We can simply not control for it, and the effect for family size found will also have the effect of increasing the chance of an absent father. One would imagine that both of them work in the same direction.

Another bad control might be the income of the household, as the number of children might influence it. But, by restricting the definition of children to only the ones that arent working, this effect should be minimized. Still, there is an effect of more children impliying in lesse worked hours. It is hard to say if the biases that income might reduce outweight new "bad control biases". The results were cheked with and without these controls, with little change, and I decided to keep the control.

The results can be seen below. There is also the visualization of the plot for the partial regression in schooling.

Table 1: Siblings Effect - with Controls

	Dependent variable:			
	BedroomsCapta	YearsEduc	BedroomsCapta	YearsEduc
	(1)	(2)	(3)	(4)
Siblings	-0.128^{***}	0.023		
	(7e-03)	(0.027)		
as.factor(Siblings)3			-0.15^{***}	0.021
			(0.01)	(0.039)
as.factor(Siblings)4			-0.275^{***}	0.055
			(0.02)	(0.077)
as.factor(Siblings)5			-0.318^{***}	0.069
			(0.036)	(0.14)
as.factor(Siblings)6			-0.359^{***}	0.054
			(0.068)	(0.262)
Observations	1,973	1,973	1,973	1,973
R^2	0.291	0.421	0.297	0.421
Adjusted R^2	0.282	0.414	0.287	0.413
Residual Std. Error	0.176	0.676	0.176	0.677
F Statistic	34.757^{***}	61.612^{***}	31.556^{***}	54.422^{***}



The probably upward biased effect on "bedrooms per capta" are all negative and significant, with a stronger effect for larger families. For Years educ, the result is statistically insignifi-

cant. This goes against the existing evidence from the literature, but is understandable given the low sample size and other problems commented in the introduction.

What about causal interpretation? With the controls, a lot of else ommited variables aren't anymore, such that defending a causal relation for the CEF is easier. Still, is the siblings variable, conditional on the controls, associated with an exogenous change? It is still hard to say. The literature brings some stories that might be relevant, such as the effect of order-of-birth in studies at the individual level, the effect of descendants that have already left the household, and the effect of the presence of the father in the household. All of those are still related to the family size, and affect the development decision.

Without a good exgoneous variation, it is always hard to defend a causal interpretation to a CEF. This is why we will consider a instrumental variable approach. It is important to say that the instrumentalization might help to get a cleaner effect, but by nature, it will actualy yield less precise estimates.

Question 9 and 10

Remark: I choose to answer both jointly, for better organization and less repetition.

In line with the same sex siblings, another instrument considered was the presence of a deficient firstborn, which might motivate the family to have more children. I don't know if such instrument is defendent in the literature.

First of all the instruments must be correlated with siblings. The evidence and discussion of this will be presented below.

Secondly, the instruments must be exogenous, this means that they both mustn't be correlated with the ommited variables in the error term, and that they mustn't affect the quality measures directly.

The fact that the instruments are related to random events, conditioned by genetics, helps, but doesn't guarantee those things. There are relevant ommited variables that are realized after the birth, that can still be correlated with the instruments, such as absence of father.

Also, there are reasons to believe that the instruments directly affect the quality measures. At least for schools, there are literatures arguing that twins have higher spillover effects than regular children.

Overall, it is not guaranteed that the instruments are exogenous, and that can't be tested. Still, lets consider the approach. There were a few instruments options considered:

- Presence of twins as last two children (1), with correlation 0.072;
- Presence of twins as last two children, at second birth (2), with correlation 0.074;
- Presence of twins as last two children, at third birth (3), with correlation 0.118;
- Presence of first two children with same sex (4), with correlation 0.021;
- Presence of first children with disability (5), with correlation 0.062.

The first instrument is the easiest to defend as being a true exogenous change in the family size, so it is the one we give more relevance to. The second and third are interesting to analise if the effect changes with the order of the twin birth, as it was done in Black et al. (2005), but, in ou scenario, we don't have enough sample to make comparisons really interesting.

The fourth and fifth ones are theoretically interesting, and nice to compare the the literature that has used them before, but are less important in the sense that require stronger assumptions to be considered exogenous.

We can see that the correlation is really low, such that the instruments can be considered to be weak. This will pose even further reduction to the statistical significance of the results.

The results can be seen below.

Table 2: Siblings Effect on Bedrooms - with Instrument

	Dependent variable:				
	BedroomsCapta				
	(1)	(2)	(3)	(4)	(5)
Siblings	8e-03	-0.198^{**}	-0.109^{*}	-0.413^{*}	-0.417
	(0.064)	(0.099)	(0.063)	(0.235)	(0.77)
Observations	1,973	1,973	1,973	1,973	1,973
R^2	0.153	0.254	0.288	-0.314	-0.33
Adjusted R^2	0.143	0.245	0.28	-0.329	-0.346
Residual Std. Error	0.193	0.181	0.177	0.24	0.241

Table 3: Siblings Effect on Schooling - with Instrument

	Dependent variable:				
	YearsEduc				
	(1)	(2)	(3)	(4)	(5)
Siblings	-0.093	-0.245	-0.115	0.267	7.829
	(0.225)	(0.379)	(0.244)	(0.676)	(14.379)
Observations	1,973	1,973	1,973	1,973	1,973
R^2	0.415	0.391	0.413	0.396	-24.758
Adjusted R^2	0.409	0.384	0.406	0.389	-25.062
Residual Std. Error	0.679	0.693	0.681	0.69	4.509

The direction of results is in line with the literature, using the cleaner exogenous variations we indeed find a negative effect related to number of siblings with most instruments. But, very few significant results, only for the super estimated measure of rooms per capta. This

was expected, the exercise already suffered with all the aforementioned sample size problems, and the IV approach only increases the variance of the estimates, specially given the weakness of the instruments.

So, not all in line with the results from Black et al. (2005), but in a similar direction.

The fourth and fifth instruments, that had a lesser quality, still recovered a negative measure for the bedrooms per capta, but not for the schooling measure.

Conclusion

Some models diagnostics are presented in the appendix. The models seem to have heteroskedasticity and one could correct the standard errors for the presented tables, with the heteroskedasticity robust standard errors. As the precision of the sample was already a problem, I chose not to do so. Also, it was studied the correlation between the variables of the models, but none presented a high VIF.

Also on the appendix, some extra exploratory analysis were made to motivate some of the controls, such as the schooling measure, and the age cuts did on the child definition.

To close up the exercise, some of the more critical changes that could've been made to this work, that might have yielded a more correct resolution, are: (i) removing the families with absent parents, and (ii) restricting the sample to more comparable families (as in question 6), even with the cost of lower sample sizes.

Appendix - Regression Results

Table 4: Linear CEFs

	Dependent variable:		
	Siblings		
	(1)	(2)	(3)
IncTot	−0e+00***	−0e+00	−1e-05***
	(0e+00)	(0e+00)	(0e+00)
IncTotCut2		−0.157***	−0.371**
		(0.043)	(0.17)
IncTotCut3		−0.178***	−0.72***
		(0.066)	(0.232)
IncTotCut4		−0.037	−0.194
		(0.148)	(0.169)
IncTot:IncTotCut2			1e-05**
			(1e-05)
IncTot:IncTotCut3			2e-05***
			(0e+00)
IncTot:IncTotCut4			1e-05***
			(0e+00)
Constant	2.456***	2.497***	2.58***
	(0.019)	(0.021)	(0.031)
Observations	1,973	1,973	1,973
R^2	9e-03	0.021	0.029
Adjusted R^2	9e-03	0.019	0.026
Residual Std. Error	0.678	0.675	0.672
F Statistic	18.474***	10.438***	8.425***

Table 5: Non-linear CEFs

	Dependent variable:	
	Siblings	
	(1)	(2)
IncTot	$-$0e+00***	$-$0e+00*
	(0e+00)	(0e+00)
I(IncTotˆ2)	0e+00***	
	(0e+00)	
IncTotLog		-0.052^{***}
		(0.014)
Constant	2.491^{***}	2.636^{***}
	(0.022)	(0.05)
Observations	1,973	1,973
R^2	0.015	0.017
Adjusted R^2	0.014	0.016
Residual Std. Error	0.676	0.676
F Statistic	14.503^{***}	16.744^{***}

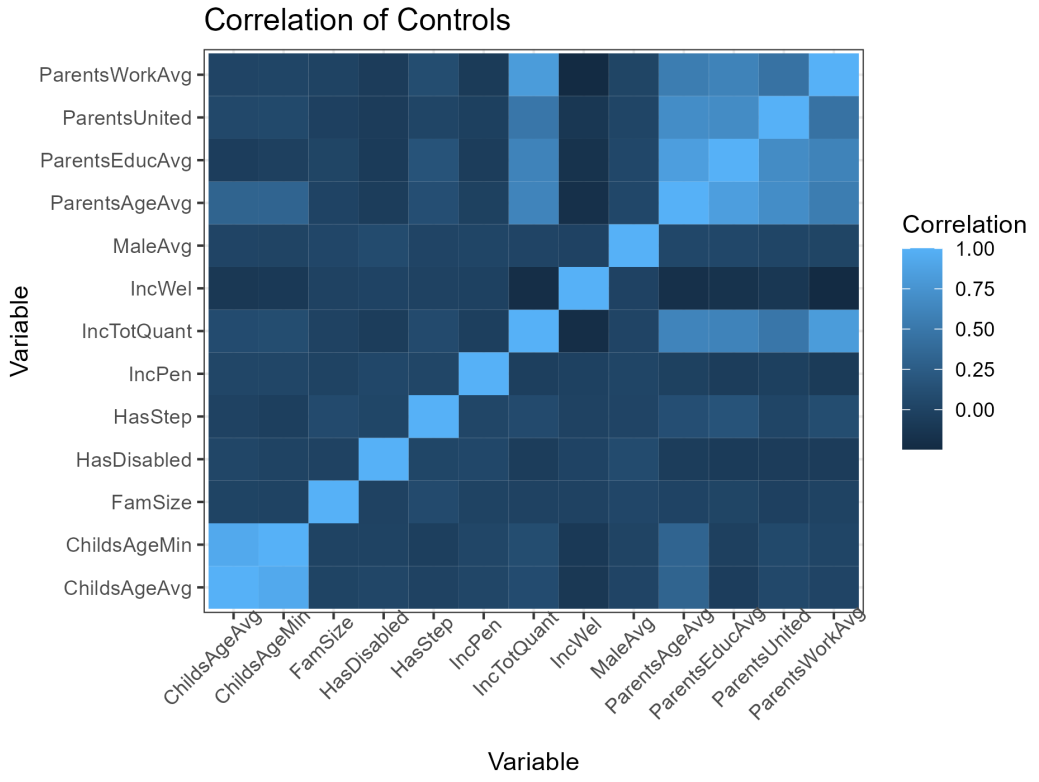
Table 6: Quantile CEFs

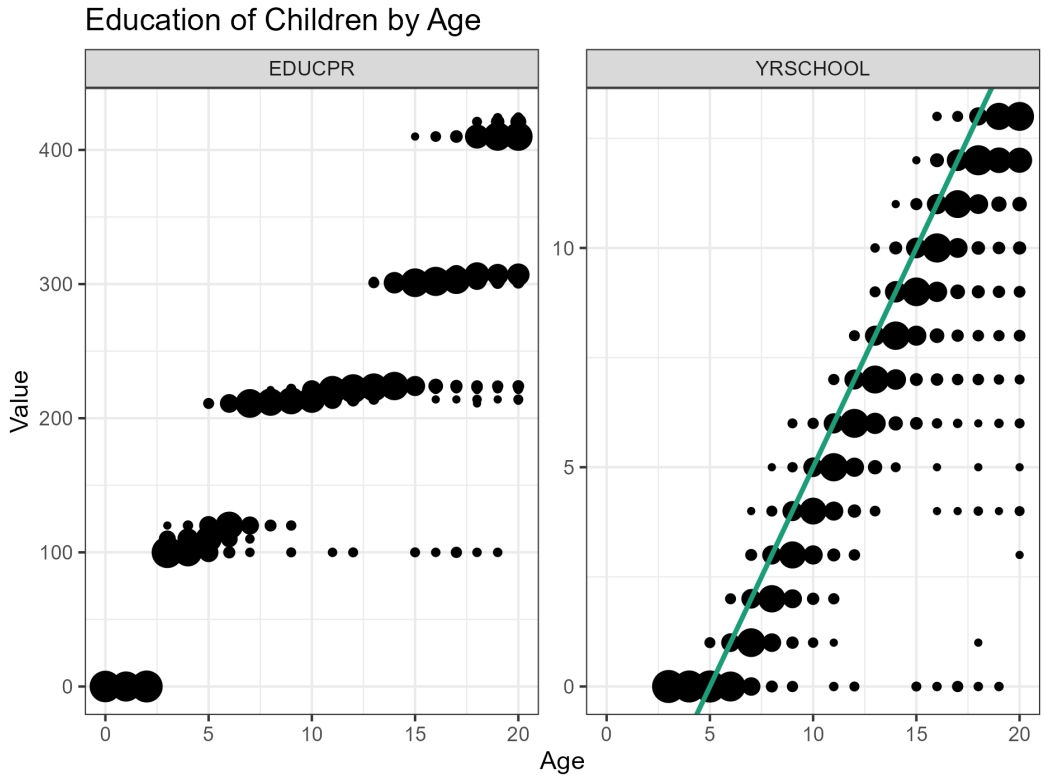
	Dependent variable:		
	Siblings		
	(1)	(2)	(3)
IncTotQuant	-0.374^{***}	-0.464^{***}	-0.73^{***}
	(0.053)	(0.125)	(0.23)
IncTotCut2		-0.514	
		(0.334)	
IncTotCut3		-1.809^{*}	
		(0.951)	
IncTotCut4		0.374	
		(3.818)	
IncTotQuant:IncTotCut2		0.747	
		(0.49)	
IncTotQuant:IncTotCut3		2.109^{*}	
		(1.088)	
IncTotQuant:IncTotCut4		-0.242	
		(3.927)	
I(IncTotQuantˆ2)			0.348
			(0.218)
Constant	2.597^{***}	2.628^{***}	2.66^{***}
	(0.031)	(0.042)	(0.05)
Observations	1,973	1,973	1,973
R^2	0.024	0.028	0.026
Adjusted R^2	0.024	0.025	0.025
Residual Std. Error	0.673	0.672	0.673
F Statistic	49.008^{***}	8.204^{***}	25.79^{***}

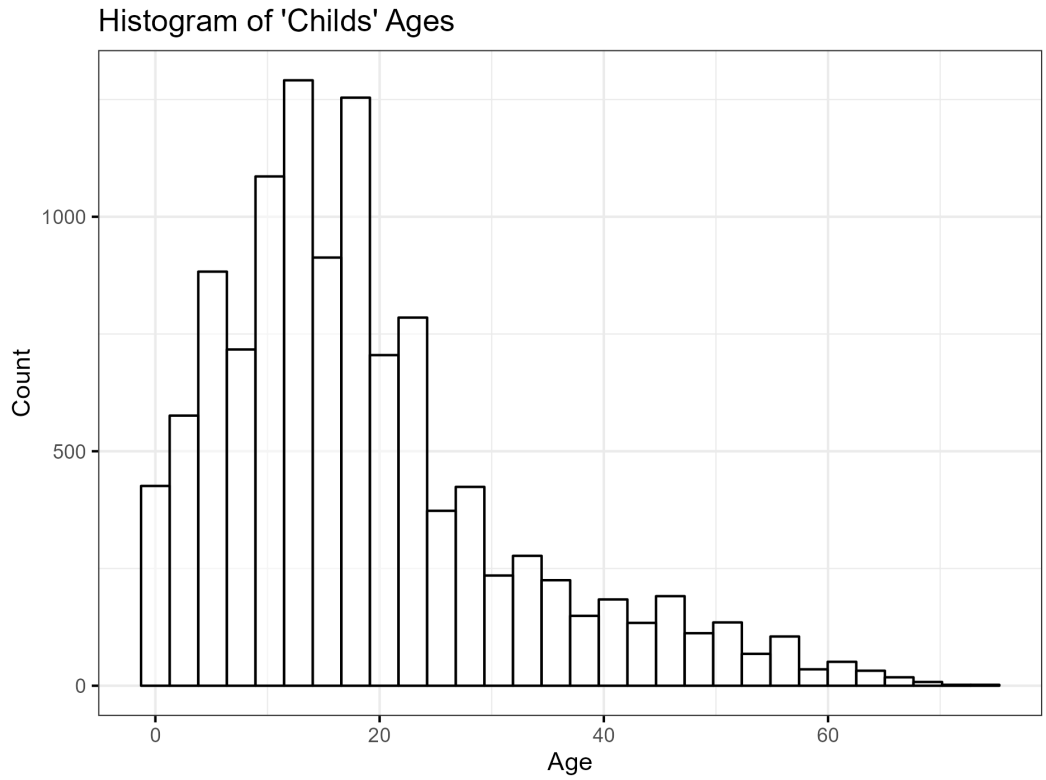
Table 7: Siblings Effect - no Controls

	Dependent variable:			
	BedroomsCapta	YearsEduc	BedroomsCapta	YearsEduc
	(1)	(2)	(3)	(4)
Siblings	-0.119^{***}	0.11^{***}		
	(6e-03)	(0.029)		
as.factor(Siblings)3			-0.149^{***}	0.076^{*}
			(0.01)	(0.046)
as.factor(Siblings)4			-0.236^{***}	0.301^{***}
			(0.02)	(0.092)
as.factor(Siblings)5			-0.276^{***}	0.341^{*}
			(0.038)	(0.174)
as.factor(Siblings)6			-0.37^{***}	0.317
			(0.072)	(0.334)
Constant	1.045^{***}	0.433^{***}	0.812^{***}	0.659^{***}
	(0.016)	(0.073)	(5e-03)	(0.024)
Observations	1,973	1,973	1,973	1,973
R^2	0.153	7e-03	0.16	8e-03
Adjusted R^2	0.153	7e-03	0.158	6e-03
Residual Std. Error	0.192	0.88	0.191	0.881
F Statistic	355.845^{***}	14.383^{***}	93.838^{***}	4.005^{***}

Appendix - Exploratory Analisys







Appendix - Model Diagnostics

Table 8:

	statistic	p.value	parameter	method
1	c(BP = 25.40)	c(BP = 3.04)	c(df = 2)	studentized Breusch-Pagan test

Table 9:

	statistic	p.value	parameter	method
1	c(BP = 3.43)	c(BP = 0.063)	c(df = 1)	studentized Breusch-Pagan test

Table 10:

	statistic	p.value	parameter	method
1	c(BP = 66.55)	c(BP = 4.08e-06)	c(df = 23)	studentized Breusch-Pagan test

Table 11:

	statistic	p.value	parameter	method
1	c(BP = 87.35)	c(BP = 1.96e-09)	c(df = 23)	studentized Breusch-Pagan test