Analysis of family income with respect to number of children in the family

Andrey Mironyuk

This paper uses current data from 2013 Annual Social and Economic (ASEC) Supplement of the Current Population Survey to examine the link between total family income and the number of children present in the family. Estimates suggest that beyond a certain point (after 4 children), the less income that family is likely to make, on the order of -$2520 per additional child, even when taking into account other explanatory variables and reweighed observations.

1. Introduction

The open nature of the March 2013 CPS allows individuals to use that data to find and search for trends. The question this paper tackles is if there is a link between family income and the number of children in the family. On the face of it, it seems natural to think that a larger family has the potential to earn more total income or that a big family implies that the parents are able to afford the numerous costs associated with children. On the other hand, it may be that, for whatever reason, bigger families make less money, and if that's the case, then this should be explored. There may be an underlying problem if, as families get bigger, they tend to have access to fewer and fewer resources. Given all of this, the research question at hand is, at what point is having an extra child correlated with lower overall income and what is the drop, quantitatively speaking?

The existing literature addresses and motivates this question to some degree. Douglas B. Downey, in his paper[1], reveals that the more children a family has, the worse the children are likely to do academically. Perhaps resources are being stretched with each additional child, which suggest that family income might not be growing along with each child. Additionally, a table by the Department of Justice[2] indicates that as "family" size increases, income increases, but the data is only available for up to a family of four, and that means, only up to 2 children, so the question of whether or not having more children is

[1] *When Bigger Is Not Better: Family Size, Parental Resources, and Children's Educational Performance*. Douglas B. Downey. American Sociological Review, Vol. 60, No. 5 (Oct., 1995), pp. 746-761

[2] http://www.justice.gov/ust/eo/bapcpa/20120501/bci_data/median_income_table.htm

correlated to total income, is still up for discussion. Although not many studies have tacked the topic of the number of actual children vs. total income, a few blogs[3][4] have, and they claim that after a certain number of kids (about 4), family income drops.

In this particular analysis, I first used summary statistics to discover if and when having additional children might be tied to lower family income, and then I ran regressions with explanatory variables to analyze this interaction.

2. Data

Main Variables

incf_all : Total family income (nominal), in dollars

nfchild : Number of family members under 18, integer, between 4 and 9 (in my restricted sample)

Explanatory Variables

state: The state the individual resided in, examples: CA, MN…

educ: A numerical score (between 1 and 5) representing the education of the individual

forborn: A dummy variable, indicating whether or not the individual was foreign born.

The summary statistics for the main variables are as follows:

Variable	Obs	Mean	Std. Dev.	Min	Max
Incf_all	202634	76965.6	90026.2	-19998	2742997
nfchild	202634	1.21561	1.355522	0	9

The approximately 12,000 individuals in the (restricted) sample and its comprehensive nature (and the fact that it confirms past studies) suggests that this sample is unbiased.

[3] http://economistsoutlook.blogs.realtor.org/2014/04/08/median-income-family-vs-household/
[4] http://www.startribune.com/local/blogs/171194091.html

3. Regression Specification

$$Y_i = B_0 + B_1 X_i + u_i \quad (1)$$

$$Y_i = B_0 + B_1 X_i + B_2 X_{2i} + \cdots + B_k X_{ki} + u_i \quad (2)$$

The basic linear regression equation in question, is summarized by (1). Other explanatory variables will be added that change the regression equation to the multiple regression form given by (2).

I will focus on the regression when the number of children in the family is 4 or more (that's when the cutoff occurs, more detail on this in the next section). My hypothesis is a negative coefficient in front of *nfchild*, and one that is significantly different than 0 (a T-test will be used to confirm this).

After the sample is weighed, I suspect there may still be problems for the OLS model. Although the regression yields a normal distribution, it's not homoscedastic (but that is touched on, and corrected in later sections)

4. Regression Results

The following 2 tables shows the mean income as a function of children

Number of Children	0	1	2
Mean Family Income	$68,078.43	$80,259.41	$89,399.25

Number of Children	3	4	5	6	7
Mean Family Income	$82,819.47	$76,871.64	$68,998.36	$62,680.58	$65,385.25

From the tables, it's clear to see there's a cutoff at 3 children, although having 7, 8, or 9 children (not all visible from the table) starts showing income start to rise slightly. Having said that, the remainder of this paper will focus on having 4 or more children to discover if there is a discernable and significant trend. The reason I am not considering 3 or more children is because, for whatever reason, the effect becomes much smaller and less significant with other explanatory variables. This is perhaps why other sources (blogs and papers) also state their results in terms of 4 children or more, so I will do the same.

Before we show the regression results, it helps to see a scatterplot of the situation, as illustrated in Fig. A below. The size of the markers indicates the frequency of that observation.

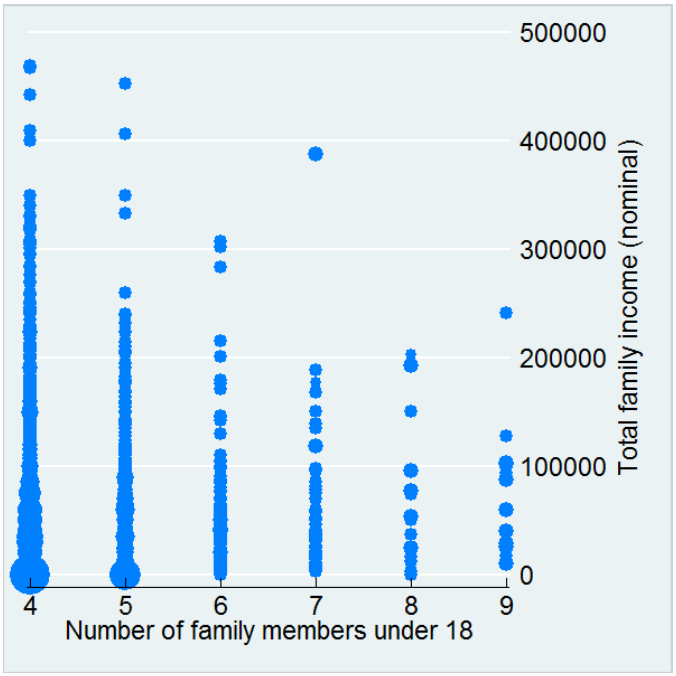

Figure A. Effect of family size on Total Family Income

The regression results are summarized in the following table. The *Short* column represent the basic regression `reg incf_all nfchild if nfchild>3 [pw = wgt]`
The *Medium* column includes the foreign born dummy and absorbs the state effect
`areg incf_all nfchild forborn if nfchild>3 [pw = wgt], absorb(state)`
Finally, the *Long* column represents the inclusion of education (which had to be absorbed due to its categorical nature, as per the suggestions). Unfortunately, I could not absorb both state and education variables, so I opted to absorb education and discard the state effect as the *Medium* regression indicated that state effects were minor anyway.
`areg incf_all nfchild forborn educ if nfchild>3 [pw = wgt], absorb(state)`

Table 1. Regression Results

	Short	Medium	Long
Number of family members under 18	-5671.4***	-5431.3***	-2520.7**
	(769.0)	(826.1)	(1176.8)
Foreign born		-17253.6***	-18839.7***
		(2381.5)	(2679.4)
Constant	96083.1***	96596.5***	90665.2***
	(4041.8)	(4304.0)	(5888.4)
Observations	11847	11847	4844
R-squared	0.004	0.041	0.115

Robust Standard errors in parentheses
* p<0.10 ** p<0.05 *** p<0.01

We can see that, before the inclusion of other variables, increasing the family size seems to decrease income on the order of $5671 with each additional child. This is the *Short* regression. However, there may be other factors at play. For instance, maybe the number of children in the family is correlated with being foreign born and following different traditions or living in a certain state (absorbed in this regression. As it turns out, these additional variables don't affect the family size effect too much, as it only drops slightly to $5,431 in the *Medium* regression. However, accounting for education in the *Long* regression (while dropping the state effect), and keeping it constant appear to drop the family size effect to $2520. Also, Stata recognized that the sample was not homoscedastic and automatically used a robust calculation for the standard errors (discussed in more depth in the next section).

Using the *Long* regression, we can test the alternative hypothesis that the actual family effect is $0, which falls outside the 95% confidence interval (with p-value 3.2%). Since the significance level for this test is the typical 5%, the null hypothesis is not rejected, although the reader can make their own judgments based on whichever significance level they feel is appropriate.

Thus, it appears that family size, is indeed tied to lower total family income (after 3 children). Maybe one explanation behind the big drop in effect and significance when the

education variable is added, could be that a large majority of the bigger families tend to have less education (and that may need to be explored in future studies), and thus, when the education level is kept constant, there is much smaller drop in the child effect.

5. Robustness Analysis

If you take a look back at Figure A (the scatterplot), you'll notice that a best fit line might be susceptible to errors. First, let's use the Short regression to look at normality. Figure B shows the results of the following commands:

```
predict resid, residual
histogram resid, normal
```

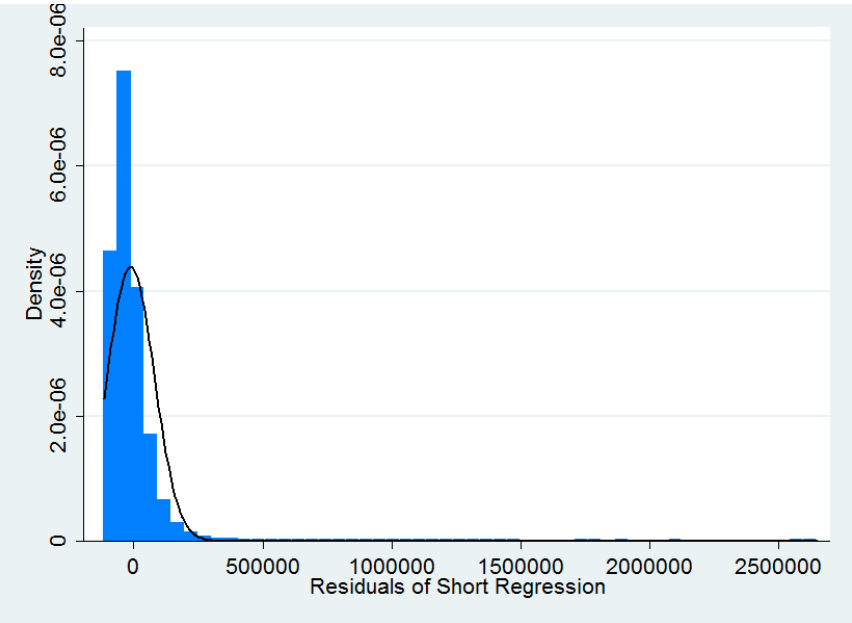

Figure B. Histogram of Short Regression

Although there are spikes, it still appears to follow a normal distribution, although this may be open to interpretation. Either way, the distribution of the residuals is not a necessary assumption for the properties we have dealt with, but it is interesting to consider. Next, we look at homoscedasticity (or lack thereof) with:

```
rvfplot
estat hottest
```

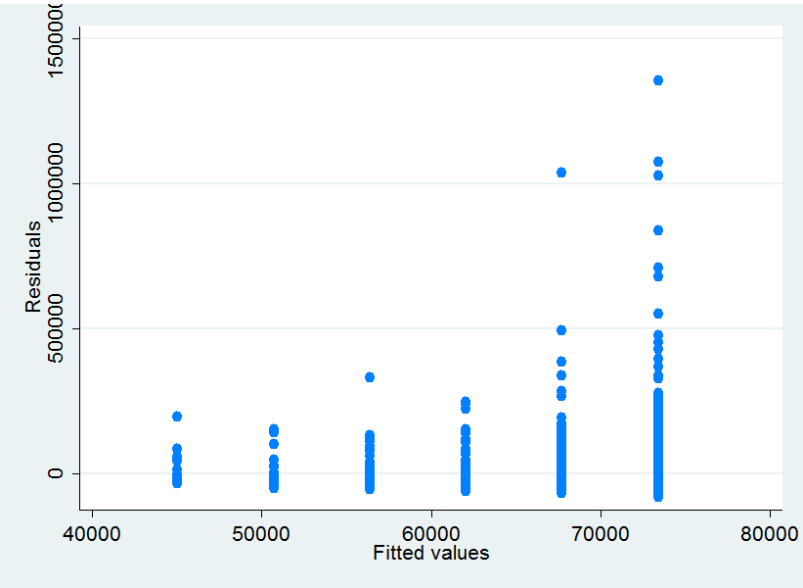

Figure C. Residuals of Short Regression

The Breusch-Pagan / Cook-Weisberg test given by the second command reveals that, indeed, the sample is not homoscedastic (as Fig. C confirms), and therefore careful care needs to be taken when interpreting the results of the regression. Correcting for this using the `robust` command in Stata doesn't affect the coefficients, but rather the standard errors. Fortunately, Stata detected this and automatically calculated robust standard errors (the ones in Table 1 above). I used those to test the hypothesis, so the results should still be valid.

Another concern is model misspecification. The scatterplot in Figure A suggests a quadratic curve fit, rather than a linear one. However, the results are easier to interpret linearly, and were still significant, so there didn't appear to be any benefit toward migrating to a different model.

Given this, overall, there appears to be a child effect on total family income (when the number of children is greater than 3). Of course, one final concern may be omitted variable bias, as the CPS data has a multitude of variables, but many are variations of a particular characteristic and thus the actual number of variables are rather limited (for instance, there's more than a 100 variables that deal with some type of income).

6. Conclusion

This paper found that there exists a child effect on total income, with each additional child (after 3, this is an important point) tied to a decrease of $5671 in total income, and $2520 when education is factored in. The results are significant, and at the same time, lend themselves to future research. Namely, is there a reason for this decline? Perhaps after a certain number of children, one of the parents is forced to stay home to take care of the children and that could account for part of the drop in income. Whatever the underlying reason, can this trend be reversed? If so, then future work may be applied in practice, to eradicate this child effect and make sure that families with more children are able to take care of them. Of course, there is also the possibility of another lurking variable that accounts for this decline, and isn't present in our dataset, but may be available in other data sets. Ultimately, future work and study is necessary in order to understand this phenomena in more detail.