**V. Empirical Methodology**

The main goal of this study is to get a better understanding of the relationship between family size and educational attainment. To achieve this, data are pulled from the Michigan State Panel Study of Income Dynamics (PSID). Here, ample data are provided on heads of households and the various areas that affect or are affected by their income. Because education and family size tend to be a product of income, plenty of data were available for both variables. The 5,416 heads of households observed provided ample amount of information relating to the research question. **Table 1: Descriptive Statistics**

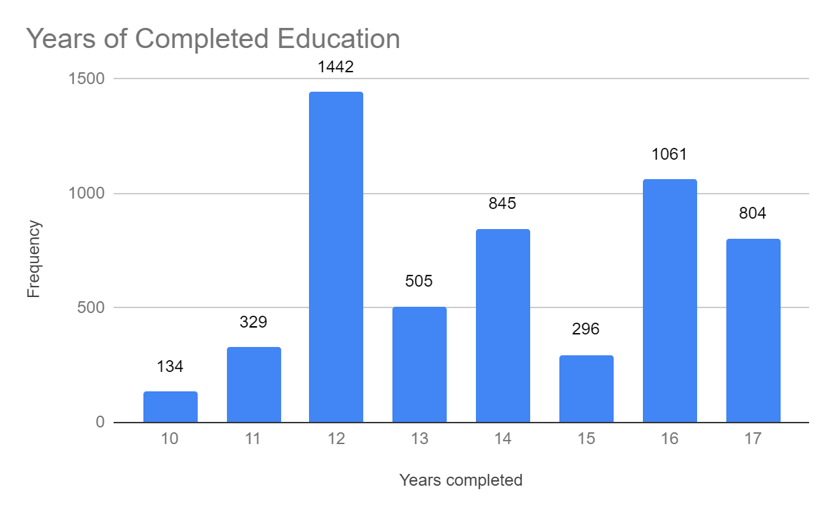
|  |  |  |  |
| --- | --- | --- | --- |
| Variable | N | Mean | Standard Deviation |
| Completed Education | 5416 | 13.98 | 2.07 |
| # of children | 5416 | 2.76 | 2.05 |
| Age | 5416 | 41.07 | 12.82 |
| Sex | 5416 | 0.71 | 0.45 |
| Rent/Own | 5416 | 0.51 | 0.50 |
| # of Cars | 5416 | 1.71 | 0.98 |
| Wages + Salary | 5416 | $53579.7 | $48351.97 |
| Metro/Non Metro | 5416 | .86 | .35 |

Table 1 shows the mean of each variable included in the study as well as the standard deviation. A dummy variable was used for sex, with a 1 representing male and 0 representing female. Looking at the chart, the statistics show that 71% of the study was male and 29% were female. While women are increasingly becoming head of households in today’s society, men still dominate that sector, leading to a male dominant study. PSID also provided data on the head of household’s age, income, as well as height. These variables provide a point of reference for household size to get an idea of how other variables affect educational attainment. Sample sizes

are similar to those analyzed in previous works within the literature review but the type of data analyzed varies somewhat. Iacovou observes data on the children at times when they are still young and in school, following them as they progress throughout their educational career (Iacovou 2008). PSID provided data on head of households who in this case, are no younger than 17 years old, though only those above the age of 22 were observed. This provides better accuracy because the research is done at an age where most individuals would have decided whether to continue their education after high school or not.

Educational attainment among the individuals surveyed turned out to be quite higher than expected. A mean of 13.98 years of education indicating around 3 years of college, with the most common number of years of completed education being 12 or 16. Interestingly, 15 years of education was the least common observation among individuals in the sample, even lower than the left and right tails. Besides this anomaly, the data are mostly normally distributed.  
**Figure 3:**

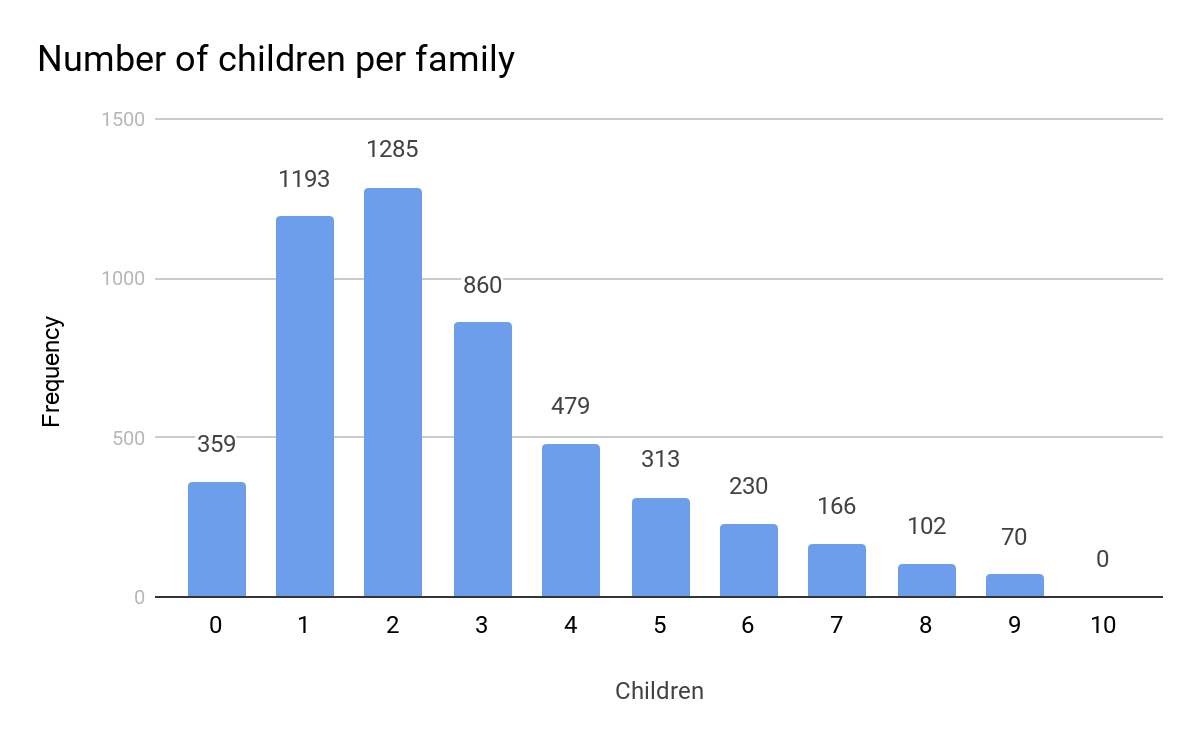
Data on household size are not normally distributed, but lie right around what would be expected when thinking of average number of children per family. The graph shows a right



skewed distribution that continues to lower after 2 children per family. Most common are families consisting of 1, 2 and 3 children, with a mean of 2.76, or about 3 children.

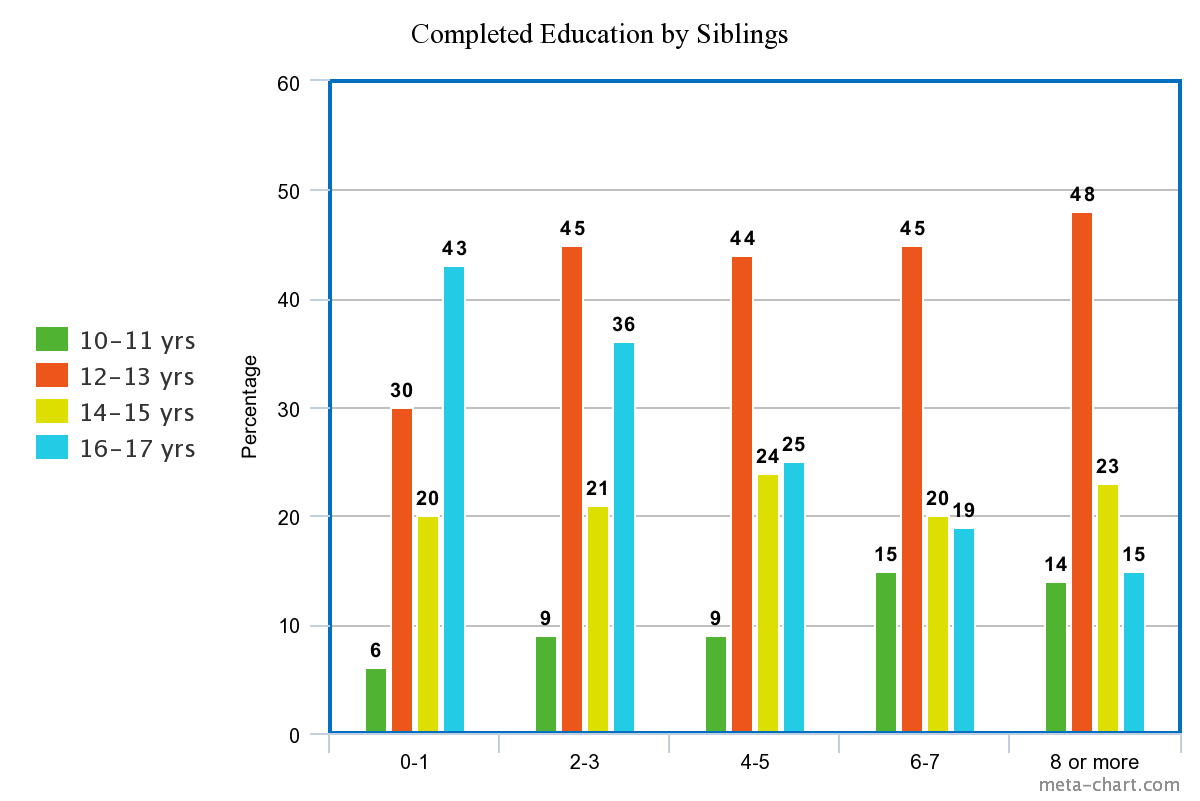
**Figure 4:**

The relationship between family size and educational attainment, is at first difficult to read. Because the US is such a developed and urbanized country, most individuals complete high school which is usually around 12 years of school. A bachelor’s degree usually puts individuals right around 16 years of schooling and so most of the data lies within the 12-16 range. A relationship does exist however. Because household size is just one of many factors that contribute to the educational attainment of individuals, a minimal effect is observed. Graph 3 below shows the percentages of educational attainment achievement per age group. Because



educational attainment is clustered within a small range of values (12-16 years), it is difficult to see any significant relationship. However, focusing on the overall trend of years of education as number of children increases, it is clear that a pattern persists. Most notable in the 10-11 age group (shown in green) and the 16-17 age group (shown in blue), as number of children increases, the percentage of individuals attaining high levels of education decreases, whereas the percentage of those attaining lower levels of education increases.

**Figure 5:**



The variables used in this study help to collectively explain what factors go into the educational attainment of individuals. Running the regression as a function of these variables will look like the equation below:  
Comped = f(#ofchildren, agehd, sexhd, ownrent, vehicles, wagessalary, metrononmet)

Rent/own, number of cars, and wages/salary are different ways to interpret income, which are expected to have a positive effect on educational attainment. Metro/nonmetro aids in understanding how the relationship between the two variables differs across different levels of urbanization. Similar to studies found in the literature review (Maralani 2008), the expected coefficient will be positive, indicating that individuals living in urban areas attain higher levels of education due to their availability of resources and proximity to schools. Looking at age, it is reasonable to expect that as age increases, educational attainment will also increase due to the mere fact that if one is older they have more time to get an education, however, because only those individuals older than 22 are analyzed in this study, an insignificant coefficient is likely to persist. All together, a positive but minimal coefficient is expected for # of children, mostly because the years of education observations lie within such a small range of values.

**VI. Results**

Running the logistic regression revealed that both *# of children* and *income* affect educational educational attainment in the expected manner. Both variables show a positive coefficient and hold t-values that are statistically significant. With all variables included in the regression, *# of children* has coefficient of (-.140). This shows that as the number of children in a family unit increases, the potential educational attainment of those children decreases. Income, also statistically significant, has a positive coefficient, proving that income and educational attainment are positively correlated. As income increases, more resources are available to

support the education of children, and more are available to divide between siblings. However, as more children are brought into the family, resources must be spread out between more children, leading to less investment per child. Table 3 shows the results for three different regressions, in which *# of children* and *income^2* respectively, are omitted to emphasize the representation of the variables in the regressions.

**Table 3: Regression results**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Regression 1 [0.208] | Regression 2 [0.1903] | Regression 3[0.093] |
| **# of Children** | **-.140 (-12.52)** | **OMITTED** | **-.193 (-15.07)** |
| Age | -.109 (-7.96) | -.114 (-8.16) | -.021 (-2.94) |
| Age^2 | .001 (7.65) | .00116 (7.47) | .000380 (1.99) |
| Sex | -.551 (-9.40) | -.517 (-8.75) | -.298 (-4.43) |
| Own/Rent | .361 (5.80) | .390 (6.24) | .738 (11.73) |
| # of Vehicles | -.036 (-1.19) | .039 (-1.28) | .074 (2.40) |
| Income | .0000831 (11.79) | .0000842 (0.76) | OMITTED |
| ln(income) | 1.090 (14.55) | 1.15 (15.43) | OMITTED |
| Metro/NonMetro | .697 (9.63) | .681 (9.33) | .947 (13.11) |
| \_cons | 4.74 (6.31) | 3.840 (5.11) | 13.224 (46.19) |

Omitting *# of children* from the regression drops the R squared value minimally but remains statistically significant. This shows that *# of children* does affect educational attainment, but is not a key determinant of the variation within the data. *Ln(income)* however, when omitted from

the regression, drops the R squared significantly while also remaining statistically significant. Income therefore, proves to be a key factor in the determination of educational attainment.

A correlation matrix including all variables ensures that multicollinearity is not present in the dataset that could potentially alter the results. Within the matrix, the highest correlation holds a healthy value of (0.44), seen between *# of vehicles* and *own/rent.* The correlation shows the logical relationship between the two variables where, assuming those who own tend to have higher income, they would also be more likely to own more vehicles. The only other correlation worth noting is the relationship between *age* and *own/rent*. The value of (0.42) however, simply explains how individuals tend to own homes as they get older and perhaps earn higher income. The results of this test are found below in table 4.

**Table 4: Correlation matrix**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Comp. Ed.** | **# of children** | **Age** | **Sex** | **Own/ Rent** | **Vehicles** | **Wages/ Sal** | **Met Non Met** |
| **Comp. Ed.** | 1 |  |  |  |  |  |  |  |
| **# of children** | -0.19 | 1 |  |  |  |  |  |  |
| **Age** | 0.03 | 0.19 | 1 |  |  |  |  |  |
| **Sex** | 0.01 | -0.10 | 0.03 | 1 |  |  |  |  |
| **Own/ Rent** | 0.17 | -0.02 | **0.42** | 0.23 | 1 |  |  |  |
| **Vehicles** | 0.08 | -0.03 | 0.25 | 0.36 | **0.44** | 1 |  |  |
| **Wages/ Sal** | 0.34 | -0.13 | 0.17 | 0.23 | 0.31 | 0.26 | 1 |  |
| **Met Non Met** | 0.14 | 0.01 | -0.01 | -0.06 | -0.08 | -0.10 | 0.08 | 1 |

Another potential pitfall for the study, is the presence of heteroskedasticity. To ensure this does not hinder the results, *estat hettest* is used. Here, prob > chi2 = 0.000, suggesting the presence of heteroskedasticity in the dataset. To correct for this, regressions thereafter include the command *vce(robust)* in order to obtain the robust standard errors. This affects the values for (t) in the regression results,however, none of the variables switch from significant to insignificant as a result of the change.