Replication of “The Primary Parental Investment in Children in the Contemporary USA is Education”

Testing the Trivers-Willard Hypothesis of Parental Investment

Logan Laszewski

2024-01-20

Address for correspondence: Logan Laszewski, Elon University
Email: llaszewski@elon.edu

The author would like to thank Dr. Nicholas Bussberg for his guidance throughout this project

Abstract

Hopcroft and Martin tested the Trivers-Willard hypothesis that high-status individuals will invest more in sons and low-status individuals will invest more in daughters using the 2000–2010 General Social Survey data. (GSS) They completed the study by examining two measures of "investments": children's years of education and their highest degree attained. The results showed that sons of high-status fathers receive more years of education and higher degrees than daughters, whereas low-status fathers receive more years of education and higher degrees than sons.

This paper is my replication of the paper written by Hopcroft and Martin, "The Primary Parental Investment in Children in the Contemporary USA is Education." The main goal is to get as close as possible to the true results they found from their analysis. Based on the information the authors give, I used the GSS data set to generate very similar results.

1 Introduction

The Trivers-Williard hypothesis that motivates this study was made in 1972, predicting that individuals in good standing will decide to invest more in their male offspring, while individuals in bad standing will choose to invest in female offspring. Statistically, in the U.S., lower-income men are less likely to get married and have children. So, the hypothesis states that it is in the interest of individuals to invest in their offspring based on sex, given these circumstances.

There have been multiple studies attacking this hypothesis, to learn more and try to back it up, especially now in the 21st century. So far, there have been mixed results, differing in methodology to complete the study and get appropriate results. There has also been the thought by Hopcroft and Martin that many reports are subject to social desirability biases since they ask questions like, “How much time do you spend with your kids?” Hopcroft and Martin questioned these studies and wanted to learn from what they did and improve on it. First, they had to figure out how to avoid bias and other issues. Then ask, what is the most effective way to track how much is being “invested” in offspring?

Hopcroft and Martin went through a deep critical thinking process to answer this question and complete the analysis in the most practical way possible. They used the GSS data set, a very reliable data set and has been collected since 1972. There are thousands of variables in this data set, plenty of which would be helpful for their research.

The primary investment parents can make for their children is in their education, in Hopcroft and Martin’s opinion. This logic comes from the fact that education is strongly correlated with economic success, health, etc., especially in the U.S. Using this claim, the testing of the Trivers-Willard hypothesis was measured by looking at the amount of education a child received. The analysis was then done using a regression model where the education variable in the GSS data was the response variable. This way, they could observe the effect of multiple predictor variables and how they influence years of education completed.

The final decision to be made was the method to measure someone's financial status. The choice was the socioeconomic index (SEI), which measures the average education, income, and social prestige associated with someone's occupation.

Based on all the information above, Hopcroft and Martin had their plan for the study in place. The regression models looking at parents' investment in their children based on years of education, led to multiple findings. The interaction of sex by father's status was significant, suggesting sons of high-status fathers do attain higher degrees than daughters. This was the case in the regression models of both Tables 3 and 4 that I replicated from the source article. Another interesting finding is that the number of siblings is negatively associated with the highest degree obtained.

2 Sample and Variable

2.1 Sample

The first step in replicating the sample used in this article was to download the GSS data set. This has thousands of variables and observations, so I needed to narrow this down to best match the numbers represented by their work. Once narrowing down my sample, I will be able to compare my values to those from the source article. The resulting observations in the article were N = 11,857 [llaszewski-ReplProj-MetaDataV1.R LAL 2024-01-17] My goal was to get as close to hitting this mark as possible. I carefully read the article and followed their thought process one step at a time. I ended with an observation amount of N = 12,210. [llaszewski-ReplProj-MetaDataV1.R LAL 2024-01-17]

Therefore, I'm certain the authors removed missing values, which included ".i," ".n," ".d," and ".a." All these missing values are denoted in different ways in the GSS data set. Since the subset used in the article is very large, getting super close to the correct dataset size would have been difficult. It is hard to pinpoint where I went wrong in my sample, but only 353 observations off is a good effort.

Given that my SEI variable was the furthest off when comparing descriptive statistics, this is one possibility of where I went wrong. My mean was 45.41 compared to the articles' 47.10. [Ilaszewski-ReplProj-DescStats_MeanSDV1.R LAL 2024-01-18] Hopcroft and Martin used the father's SEI when the respondent was 16 years old by asking the respondents about their father's job when they were that age. They could have possibly eliminated more observations if they couldn't calculate a new SEI based on responses to this process. This would lead to my values for that variable being off and my total observation count being a little higher than it should be. Other than that variable, my numbers were very accurate, so I was confident I made a nearly correct sample.

2.2 Variables

To match my variables with theirs in the article, I also referred to the descriptive statistics table. (Table 1) As listed in the GSS dataset, the variables I needed were: educ, sex, pasei10, age, race, and sibs. In the table, there were also frequencies listed with five different possible outcomes. By observing, this was for the highest degree achieved by the respondent: less than high school, high school, associate degree, bachelor's degree, and graduate degree. After looking through the codebook, I was able to determine this must be coming from the degree variable, representing respondents' highest degree. 0 = <HS, 1 = HS, 2 = Associates, 3 = Bachelor's, 4 = Graduates. So, by getting the frequency of each option, I could replicate the descriptive statistics, adding

my seventh variable to my sample. [llaszewski-ReplProj-DescStatsFrequenciesV1.R LAL 2024-01-18] I also had to reference the variable year, just so I could subset the data only from 2000–2010. [llaszewski-ReplProj-subsetyearsV1.R LAL 2024-01-18] I decided to keep most of these variables as they appeared in the GSS data set. I made this decision since there were not too many variables I was dealing with, and I could easily distinguish one from another the way they were. Also, I liked the simplicity of each variable name, and none were annoying to type.

It's also important to look for any variables the article might have recoded for simplicity in their study. The first variable I had to recode was race. Using the descriptive statistics table (Table 1) in the article, I saw race (white = 1). The GSS data set does have White coded as 1, but also has Black = 2 and Other = 3. I observed the mean race as 0.81 in the article. Since this was the case, I knew that if you weren't white, every other race was being coded as 0. So, in my script, I recoded it so that White = 1 and Other = 0. [llaszewski-ReplProj-VariablesV1.R LAL 2024-01-15] Since I adjusted the source variable, I renamed my newly coded variable race_updated.

Afterward, I went through a very similar process with the sex variable. This was labeled as sex (male =1). In the GSS data set, male = 1, but female = 2. All I needed to do was use an ifelse function that made all 2's become 0's. This was a very straightforward process and pretty similar to the race variable procedure. I once again renamed the variable I adjusted; this was now called sex_updated.

I completely removed the source variables race and sex from my sample. [llaszewski-ReplProj-VariablesV1.R LAL 2024-01-15] Since I would not need them for my analysis, I don’t want to accidentally confuse them with each other. This way, I was able to simplify my sample and still only have the seven variables that I needed.

It is also necessary to look through your observations for values that look suspicious. Even after removing missing values, sometimes there could be values that don’t make sense for a specific variable. I used histograms to examine the dataset for outliers or unexpected values. This was straightforward for most variables. I can look at if age, educ (0-20), socioeconomic index (1-100), etc all fall within reasonable ranges. The only variable that ended up being concerning to me was the number of siblings. There were some very high observations (25+). I left them in the data set since I still had almost identical mean and standard deviations as in the article.

3 Verification of Analysis

I will review my analysis compared to the author's analysis in the source article. The analysis the article looks at uses multiple different regression models. The regression models look at the relationship between years of education (educ) and several other predictor variables. The objective of this model is to understand how these different variables—race, sex, etc.—contribute to variations in the respondent's number of years of education completed.

There are three important pieces to compare when observing whether my analysis was accurate to those in the source article. The first piece is looking at the coefficients for each variable, specifically if they are the correct sign (+/-). This coefficient implies how the years of education move when holding all other variables, and there's a one-unit increase in the observed variable. So, if this number is similar, then the models are both predicting the same change. Next is the standard error, which is in parenthesis. They indicate the precision or uncertainty associated with the estimated coefficients. Then there is the p-value, which holds a lot of importance. There is a scale from one to three stars, representing how significant the p-values are. This means how much evidence there is that a change in the variable constitutes the change in years of education completed.

There are four models I replicated from the source article. First was Table 3, which looks at all cases. Model 1 is without siblings as a predictor variable, and Model 2 includes siblings as a predictor variable. For Model 1, all the coefficients have the same sign. Along with almost identical standard errors. All my p-values were significant when the source articles were. [llaszewski-ReplProj-RegressionTable3V1.R LAL 2024-01-19]

However, my sex variable was significant when the article had a p-value greater than 05. Model 2, which adds to the number of siblings as a predictor variable, also had very similar results. [llaszewski-ReplProj-RegressionTable3V1.R LAL 2024-01-19] The only thing that sticks out is the sex and father's SEI interaction and the number of siblings' p-values. The p-value for sex and father's SEI was found to be <0.001 in my sample, but in the source article, it was only <0.05, meaning mine was more significant. The siblings' p-value was the exact opposite, and my sample was less significant than theirs. [llaszewski-ReplProj-RegressionTable3V1.R LAL 2024-01-19] But, since both are still significant when they should be, it is still a similar trend and isn't an issue. Finally, when looking at the R^2 values for each model, they were both exactly .09 off. [llaszewski-ReplProj-RegressionTable3V1.R LAL 2024-01-19] The articles R^2 both are around 0.2. This is a low R^2 and suggests that the regression model explains a relatively small proportion of the variability in the number of years of education completed.

The second table I replicated was Table 4. I decided to include this table to further verify my sample and ensure my analysis has similar results. Table 4 no longer looks at all cases; it only includes the respondents who are over 24 years old. This created a new sample, which in the source article was N = 10,857. My new sample was similar, sitting at N = 11,187. [Ilaszewski-ReplProj-RegressionTable4V1.R LAL 2024-01-19] The table layout was the same, with no change in the predictor variables for Models 1 and 2.

Since Table 3 was similar and there wasn't much of a change in my sample, I once again expected similar results. I was right, as Model 1 had coefficients with all the same signs and almost identical standard errors. The p-value was a little more significant for the sex variable than the source articles, just like in Table 3. [Ilaszewski-ReplProj-RegressionTable4V1.R LAL 2024-01-19] Model 2, including the number of siblings, had correct signs for coefficients and great standard errors. All p-values were nearly exact, with my sex and father's SEI interaction being slightly less significant. [Ilaszewski-ReplProj-RegressionTable4V1.R LAL 2024-01-19] Once again, the R^2 values are just slightly off for each model. Both have increased a little but remain low.

Overall, Tables 3 and 4 don't vary much when looking at the analysis of both. Only looking at ages greater than 24, cut the observations by nearly 700, but resulted in almost the same results.

To conclude, my results were very accurate and closely resembled the source article’s findings. The main difference was within the father's SEI and sibling's interactions. Specifically, in Table 3, Model 2. [llaszewski-ReplProj-RegressionTable3V1.R LAL 2024-01-19] This most likely goes back to the fact that my mean and standard deviation were the furthest off for the father's SEI variable. Due to that, it isn’t surprising that I would see slightly different results when it comes to the analysis.

With the sibling's variable, it is possible that they did remove some outliers. Since there were a decent number of respondents with 25+ siblings, which seemed a little out of the ordinary, as I mentioned earlier, my p-values were significant when the source articles were, and nothing else stood out as being off.

4 Conclusions

In all, my replicative results did a great job of representing those in the source article. I learned that it may take a lot of trial and error to produce results in a study. Having a goal and target for the results I want to reach made this process a lot easier, and I understand why real studies are difficult and can take years to complete.

In my efforts to replicate these results, the most difficult task was to pick the brains of the authors and try to understand exactly what they were trying to accomplish and why. Most of this process occurs in the data management stager, which was very time-consuming but also extremely important. The actual analysis is exciting and didn’t take much time at all to complete.

Even though my results aren't an exact repetition of the study Hopcroft and Martin completed, I was able to get close. Using my sample, it is possible to conclude the same results that they found in their work. It would be interesting to do more research and try to improve on this study which questions the Trivers-Williard hypothesis.

References

[1] Hopcroft, Rosemary L., and David O. Martin. "The Primary Parental Investment in Children in the Contemporary USA Is Education - Human Nature." SpringerLink, Springer US, 8 Mar. 2014, link.springer.com/article/10.1007/s12110-014-9197-0.