Date: April 23, 2019
To: Leana Wen, President, Planned Parenthood
From: Julian Vazquez, Junior Data Analyst, Janzen Consulting Group
RE: Effect of years of education in number of children in the US

According to a report from the Census Bureau, individuals with a higher level of education have fewer children than those with a lower level of education[1]. This is significant, because with more people going to college than ever before, this relationship between education and childbearing could potentially change the traditional family structure in the United States.

Fertility in the United States has been steadily decreasing in the past few decades[2], and education may be playing a role in this trend. I hypothesize that the education level of an individual can help explain the number of children they end up having. To evaluate if this is true, I will focus my efforts on studying adults in the United States. Preliminary findings show that education level can indeed help predict the number of children someone has.

The data for the following analysis comes from the General Social Survey (GSS) dataset, a well-respected survey that provides policy makers with a clear and unbiased perspective of what Americans think and feel about a wide range of issues[3]. The study will exclusively focus on two variables: education level and number of children. Education level, defined by the GSS as years of education, ranges from 0 to 20, showing spikes in the 12 (high school) and 16 (college) units. Number of children ranges from 0 to "Eight or more", which was redefined as simply "8" for analysis purposes. Very few observations fell in this category, therefore it should not have a significant effect on the findings.

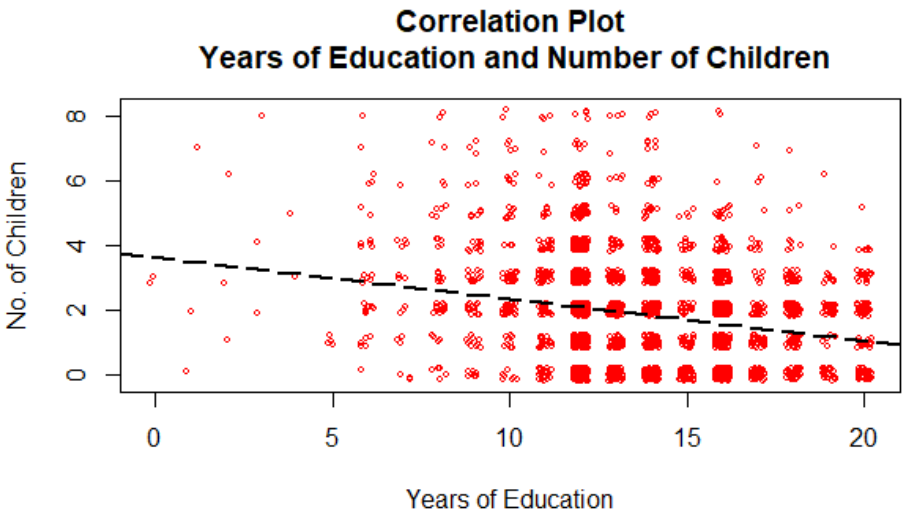
Table 1: Summary Statistics

	Freq	Min	Max	Med.	Mean	SD
Education	2858	0.0	20.0	13.0	13.7	3.0
No. of Children	2859	0.0	8.0	2.0	1.9	1.7

As previously stated, the variables used for this regression model are years of education and number of children. Single linear regression analysis was used to test if years of education significantly predicted number of children, with an alpha of 0.05. The results of the regression indicated that the predictor explained 5.3% of the variance ($R^2 = .053$, $F(1, 2851) = 160$, $p < .001$. Given our alpha, the

[1] Wetzstein, C. (2011). Education level inversely related to childbearing. The Washington Times. Retrieved from https://www.washingtontimes.com/news/2011/may/9/education-level-inversely-related-to-childbearing/
[2] Chappell, B. (2019). U.S. Births Dip To 30-Year Low; Fertility Rate Sinks Further Below Replacement Level. NPR. Retrieved from https://www.npr.org/sections/thetwo-way/2018/05/17/611898421/u-s-births-falls-to-30-year-low-sending-fertility-rate-to-a-record-low
[3] About the GSS (n.d.). The General Social Survey. Retrieved from http://gss.norc.org/

results were extremely statistically significant. Since we only have one independent variable (IV), this means that both our model and our IV are statistically significant, and we can reject the null hypothesis stating that education level would not help predict number of children. Our coefficient is meaningful and estimates that each additional year of education corresponds to an average decrease of 0.13 in number of children. In other words, every 7.7 additional years of education correspond to one less child. This would mean that as the base level of education increases, individuals can be expected to have fewer children. Because our data comes from the GSS dataset and is representative of adults living in the US, we are able to generalize our findings to the population.

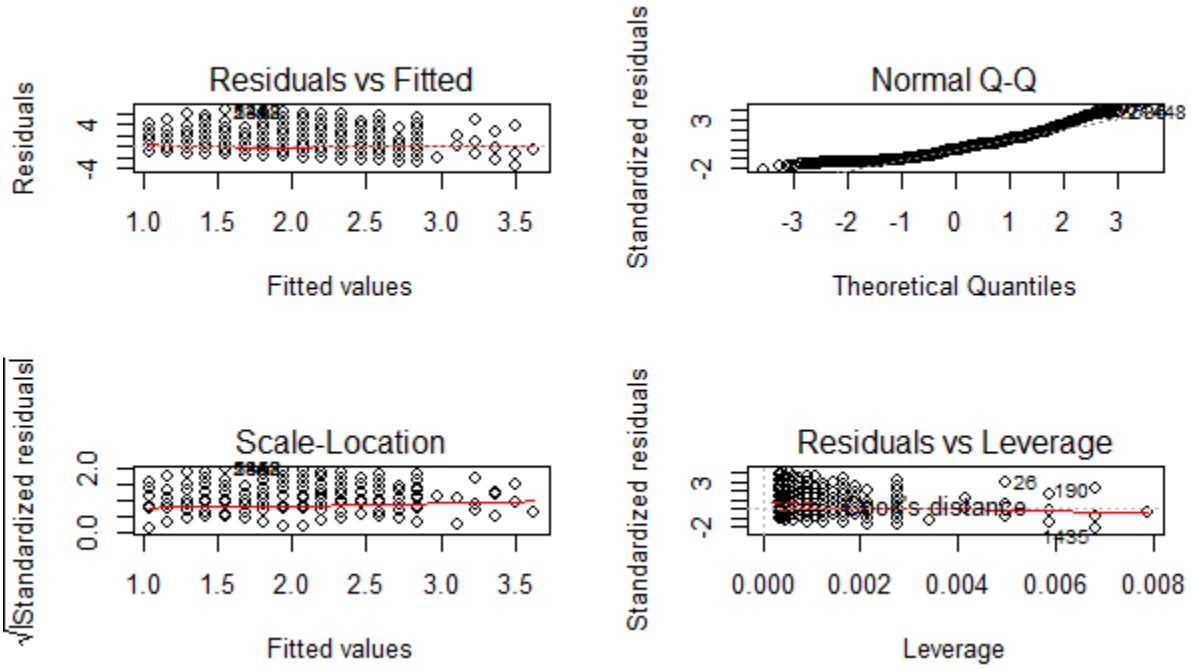


$$No.\,of\ Children = 3.63 - \ 0.13 * Years\ of\ Education + \ \varepsilon$$

One potential weakness of this study is the fact that it does not include individuals who reported having more than eight children. However, since this is such a small portion of the population, I do not expect my results to have deviated significantly if I had included this data. The same applies to the fact that the data set caps at 20 years of education, since very few people go beyond that.

$$No.\,of\ Children = 3.63 - \ 0.13 * Years\ of\ Education + \ \varepsilon$$

Diagnostic plots



Residuals vs Fitted: flat, straight line and no abnormal patterns. Linearity is ok, and no homoscedasticity issues

Normal Q-Q: deviations from the line at both tails, but tracking along nicely in the middle

Scale-Location: relatively equal spread of observations, no homoscedasticity issues

Residuals vs Leverage: no red dashed lines, so no influential outliers

<u>Closer look at possible violation of normally distributed errors (QQ plot)</u>