The Relationship Between One's Life Background and Economic Outcomes in the Future

Peter Chapman

The University of Delaware

May 15, 2025

Github.com/petecht66

ECON306: Data Analysis for Economics and Business

Analysis of the Paper

Introduction to the Paper

A paper by Eric Chyn and Lawrence F. Katz titled 'Neighborhoods Matter: Assessing the Evidence for Place Effects' attempts to isolate the causal relationship between living areas and the outcomes of adults and life courses of children. To be more specific, the authors of the paper study the causal impact of neighborhoods in the two decades since early research on the Gautreaux Assisted Housing Program, a housing program in the city of Chicago from 1976 to 1998. The focus of this study is evidence from high income countries, but the authors believe that lessons from neighborhood effects in developed countries can be utilized to understand neighborhood influences in developing countries as well.

Neighborhood Poverty Rates

The authors of this paper clarify that they used publicly available US data in analyses with the hypothesis that place of residence does matter. In order to study neighborhood effects, the authors studied 741 commuting zones: aggregations of US counties based on commuting patterns in the 1990 Census. The first analysis conducted was a series of four simple linear regressions between a commuting zone's poverty rate and various outcomes for adults and children. The poverty rates were collected from the 2000 Decennial Census, and the authors point out that these poverty rates summarize a bundle of characteristics of a living area. The independent variable for all four of these regressions is poverty rate, and the four dependent variables are adult employment rate, life expectancy, upward mobility, and mean test score. It is important to note that these are four different bivariate regressions with just one of these dependent variables being studied at a time. Figure 1 in the paper shows the results of these four simple linear regressions. Using these results, the authors conclude that there is a strong association between an area's poverty rate and these four outcomes. Being more specific, employment rate, life expectancy, upward mobility, and mean test score for children are all predicted to decrease as poverty rate increases. Furthermore, the authors claim that Table 1 shows that all four of these relationships in Figure 1 are statistically significant at the 1% confidence level. Although this table does not include confidence intervals for the impact of the individual predictors, the R-squared for these simple linear regressions is provided.

Adult and Child Outcomes in Chicago

The authors studied poverty rate, adult employment rate, and upward mobility even further within the city of Chicago. Shown in Figure 2, the authors created heat maps for these variables to illustrate within-city patterns using Chicago as an example. Dark red indicates areas with poor performance in these metrics, dark blue indicates exceptional marks, and yellow

indicates average marks. With these heat maps, the authors are making comparisons between commuting zones in the sample in regards to these performance metrics. Panel A, which graphs poverty rate, shows decent variability within the city. Panels B and C, which graph adult employment rate and upward mobility, illustrate that high-poverty areas perform poorly in these metrics. The authors offer two possibilities for these associations. First, there is the possibility that neighborhood environments do in fact have causal impacts on adults and children. Instead of hastily offering this as a conclusion, the authors present the idea that these observed patterns could reflect non-random sorting of the people who lived in disadvantaged areas.

Neighborhood Effects for Adults and Children

At the top of page 204, the authors start to clarify that there are multiple factors that contribute to the outcomes for an individual. Despite this multifactorial nature of outcomes, the authors focus heavily on neighborhood effects, of which there are so many. These include but are not limited to school quality, access to employment, peer influences, and neighborhood safety. This long list of factors suggests that the best way to predict an individual's outcomes is to consider a multiple linear regression, where multiple predictor variables are used instead of just one. These neighborhood effects can be organized into three statistical categories. The first is endogenous effects, which is any effect on outcome that stems from one's peers. The second is exogenous effects, which are effects on outcome that stem from one's neighbors. The final category is correlated effects, which refer to the idea that certain factors in an individual's life are dependent on one another. This suggests that there might be collinearity (linear dependence) among factors considered in a multiple linear regression to predict one's outcomes.

The study then summarizes the findings of a housing voucher experiment. Chicago families were organized into three groups: those who moved to low-poverty areas, those who received Section 8 vouchers, and those who had no assistance in moving. The study found that the families in the low-poverty areas reduced poverty by 35% on average, and these families felt safer and healthier. However, there were limited improvements for these families from an employment and economic standpoint. The authors of the article claim that confounding factors such as discrimination, poor transportation, and community violence could explain this lack of economic improvement.

Researchers focused on older children who moved to low-poverty areas from this study. There was a notable difference between boys and girls studied. Female children received improvements in education and health, but male children experience mostly negative effects from the move. Past the housing voucher experiment, future studies have shown that younger children experience positive outcome effects (adult earnings, college attendance, job results, and criminal justice results) from exposure to low-poverty areas as opposed to older children. Overall, moving to these low-poverty areas seemingly presents younger children with better education, better peer influence, and better future outcomes as adults.

Bias and the Lack of Randomness

The authors of the article do not directly reference homoscedasticity or heteroscedasticity, but they do identify that the Gautreaux experiment did have some flaws in it. Katz and Chyn reference an article by Mendenhell, DeLuca, and Duncan that points out that the placement type of certain housing units (suburban or rental unit) was systematically related to other factors, which means that this was not a fully randomized experiment. Fortunately, as Katz and Chyn identify, future neighborhood-effects studies have relied on randomized field experiments in order to address the concern of selection bias.

In the discussion section, the authors explain that this selection bias also has an impact on the findings of the study. Not having randomization of housing units selected has essentially inflated the magnitude of neighborhood effects on an individual adult's economic outcomes. Notably, the authors claim that these effects are most likely somewhat still relevant; however, the true magnitude of these effects on an individual's economic outcomes are not completely accurate in many neighborhood-effects studies.

Data Analysis

Introduction to the Data Analysis

To supplement the paper by Katz and Chyn, data has been collected from Opportunity Atlas, who can be found online at opportunityatlas.org. This data set, marked by commuting zones, contains many social and economic variables that can be studied. The goal is to possibly identify any causal relationships between variables that could possibly indicate a connection between an individual's background and their outcomes in life. There are over 400 variables in this data set, but only a select few have been analyzed for this section of the paper. Variables are described as they are introduced in the analysis. To perform this statistical analysis, STATA is used through the Apporto Cloud Mounter from the University of Delaware's Lerner Desktop.

Percentage of Children to Graduate College

To begin the analysis, a dependent variable must be chosen to serve as the foundation for statistical analysis. The variable chosen for this is the percentage of children who grew up in a commuting zone who obtained a four-year college degree, pooled across both race and gender. This is a good dependent variable to analyze because it represents a strong outcome for children in a commuting zone. It is generally understood that earning a college degree helps set an individual up for success (especially financially) in their adult life. The chosen dependent variable is pooled across both race and gender for a specific reason: these characteristics (race

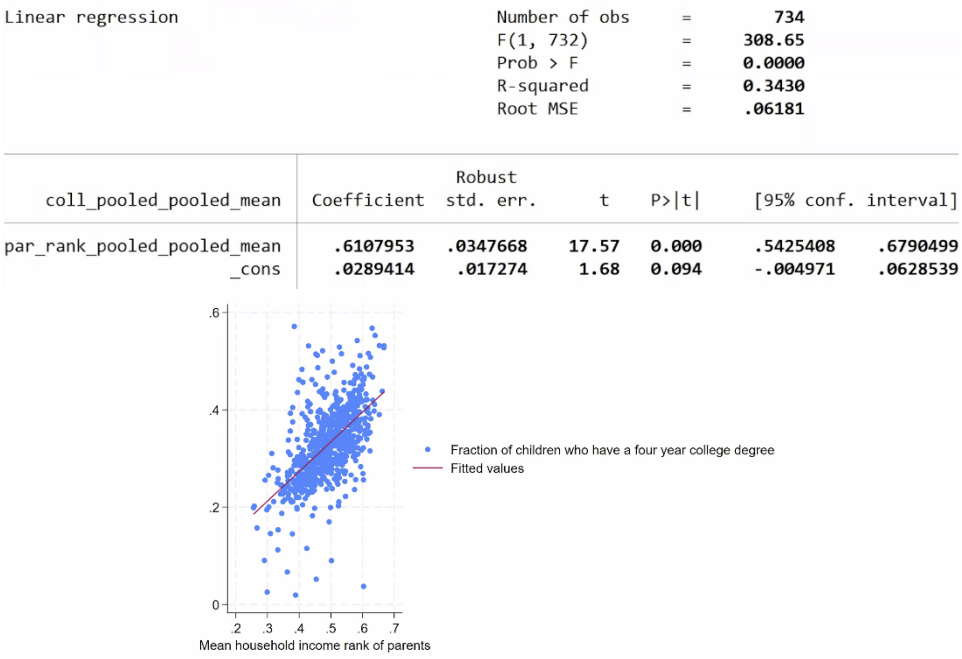
and gender) could be studied separately as predictor variables to see if they have a causal impact on this dependent variable. STATA outputs summary statistics for this chosen dependent variable, which includes 734 observations. The mean of .3276544 means that on average, about 32.8% of children in a commuting zone have a four-year college degree. The variance of 0.0058077 suggests that there is a variation of about 0.5% across the commuting zones.

Before a regression is analyzed, it is important to test the dependent variable on its own. To do this, a one sample t test has been performed on the percentage of children who grew up in a commuting zone who obtained a four-year college degree, pooled across both race and gender. The null hypothesis is that the mean is equal to 0.333, which is supposed to represent one-third. In the context of the variable, the null hypothesis is that the mean percentage of children in a commuting zone who obtain a four-year college degree is 33%. It is believed that one-third is a reasonable expectation for the fraction of children in a commuting zone who obtain a four-year college degree. The alternative hypothesis is two-sided, meaning that the mean percentage of children in a commuting zone who obtain a four-year college degree is different from 33% in either direction. A standard confidence level of 5% is used for this two-sided test. From STATA, the probability of this two-sided test from the sample data is 0.0578, which means that there is a 5.78% chance that the null hypothesis is true. This is greater than 0.05, so the null hypothesis is not rejected for this test. There is not enough evidence from the sample data to conclude that the mean percentage of children in a commuting zone who obtain a four-year college degree is different from 33%.

Bivariate Regression

With this further information about the response variable, a bivariate regression can be conducted to see if there is a causal relationship between a predictor variable and the percentage of children in a commuting zone who obtain a four-year college degree. The predictor variable chosen for this study is the mean household income rank for parents, pooled across both race and gender. For this regressor, parents are ranked relative to other parents with children in the same birth cohort. This variable is measured in percentile rank, so a value of .75 means the mean household income is in the 75th percentile. Both the independent and dependent variables are pooled across both race and gender in an attempt to isolate the relationship between mean household income for parents and the percentage of children who obtain a four-year degree. Essentially, the purpose of this bivariate regression is to investigate the relationship between parental income and a child's ability to obtain a four-year degree. It is hypothesized that there will be a statistically significant and positive association between these variables, as it is generally believed that an increase in parental income improves the percentage of children who are able to obtain a four-year degree. Robust standard errors are used for this bivariate regression.

Figure 1 below shows both the regression results table and the scatter plot of these two variables in STATA. The slope parameter for the mean household income for parents variable is 0.6107953. Because both variables represent percentages, this slope parameter can be a little bit tricky to interpret. Essentially, this suggests that if the mean household income rank for parents in a commuting zone increases by 1%, it is predicted that the percentage of children who obtain a four-year degree increases by around 0.61%. This is a decently large positive slope parameter, and STATA reports the t-test probability to be 0.000. These facts, along with the fitted line from the scatter plot, suggest that there is a statistically significant and positive correlation between mean household income rank and the mean percentage of children who obtain a four-year degree. These variables are most likely positively correlated because an increase in a parent's money allows their children a larger financial opportunity such as attending college. The R-squared value for this model is 0.343, which means 34.3% of the variability in the response variable can be explained by the lone predictor variable. This is not a very high value, so this model is certainly not perfect in explaining the mean percentage of children who obtain a four-year degree.



Linear regression

Number of obs	= 734
F(1, 732)	= 308.65
Prob > F	= 0.0000
R-squared	= 0.3430
Root MSE	= .06181

coll_pooled_pooled_mean	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
par_rank_pooled_pooled_mean	.6107953	.0347668	17.57	0.000	.5425408	.6790499
_cons	.0289414	.017274	1.68	0.094	-.004971	.0628539

Figure 1: Regression results table and scatter plot of the bivariate regression between mean household income for parents and fraction of children who have a four-year college degree.

Multivariate Regression with A Control Variable

To improve the fit of the model, and to test for omitted variable bias, a second predictor variable can be added to the regression model. Omitted variable bias is when a regression model

like this one leaves relevant explanatory factors out, and the estimated impacts of the independent variables in the model are inflated. Because there are multiple explanatory factors that go into whether or not a student obtains a four-year college degree, the true impact of parental income rank is most likely not as large as the bivariate regression above suggests. The key control variable added for this step is the percentage of children claimed by two people, which is the fraction of children who had two parents at home. This is a control variable because it is an additional factor that goes into whether or not a child obtains a four-year degree. Generally speaking, having two parents at home helps a child academically, so adding this predictor could give a better estimate of the predictor variable of interest, which is still the average household income rank for parents.

Figure 2 below shows the multiple linear regression results generated by STATA when this control variable is added. The estimate for the household income rank decreases from around 0.61 in the bivariate regression to 0.5555 when the control variable is added. This is not much of a decrease, but it suggests that there was a little bit of omitted variable bias in the bivariate regression. The control variable, the percentage of children who had two parents, is considered statistically significant by STATA, with a slope parameter of 0.1181553 and a p-value of 0.000. This suggests that there is a strong positive correlation between having two parents at home and children earning a four-year college degree. The R-squared value also slightly increased to 35.92%, which still isn't the strongest explanation of variability. From these results, it is reasonable to conclude that both variables are statistically significant in explaining the percentage of children who receive a four-year college degree, even in the presence of the other predictor variable. In the context of the control variable, it is reasonable to say that having two parents at home slightly increases the probability of a child earning a four-year degree. This makes sense, as there is increased academic support and financial opportunity.

Linear regression

Number of obs	= 734
F(2, 731)	= 212.85
Prob > F	= 0.0000
R-squared	= 0.3592
Root MSE	= .06109

coll_pooled_pooled_mean	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
par_rank_pooled_pooled_mean	.5555833	.0434623	12.78	0.000	.4702575	.6409091
two_par_pooled_pooled_mean	.1181553	.0334828	3.53	0.000	.0524214	.1838892
_cons	-.0342105	.0192092	-1.78	0.075	-.0719223	.0035013

Figure 2: Multiple linear regression results when the first control variable is added to the model. Generated by STATA.

Adding a Third Predictor Variable

To improve the model's R-squared and overall fit as well, a second control variable can be added to the model. This new predictor variable is the mean percentage of females who claimed that they had a child between the ages of 13 and 19. This predictor is certainly relevant, as having a child between these ages most likely makes it difficult to complete a four-year college degree. Therefore, it is predicted that there is a strong negative relationship between this predictor variable and the response variable, the percentage of children who obtain a four-year college degree. Although this hypothesis exists, this new predictor is still just a control variable. Robust standard errors are used for this analysis, just as in previous ones.

Figure 3 shows the results from this multivariate regression analysis with three predictor variables included. All three variables are considered statistically significant, even in the presence of the other regressors. Notably, the estimate for the mean household income rank for parents variable decreases from around 0.555 to 0.257. This suggests that if the mean household income rank for parents in a commuting zone increases by 1%, it is predicted that the percentage of children who obtain a four-year degree increases by around 0.257%, holding the other predictor variables constant. In addition, the estimate for the two parents at home predictor variable changed from around 0.11 to -0.28. Both of these are significant changes, which suggests that the previous regression model was suffering from omitted variable bias. Bizarrely, the estimate for the two parents at home predictor is now negative, which seems wrong in a real world sense. The estimate for a female who had a child between the ages of 13 and 19 is about -.744. This is a decently large negative value, which confirms the suspicion of a strong negative correlation. This makes sense, as it becomes more difficult for a female to complete a four-year degree when they have to take care of a child starting in this young age range. Notably, the R-squared jumps up from 35.92% to 50.25%, which means this new multivariate model does a significantly better job of explaining the variability in the percentage of children who obtain a four-year degree.

Linear regression

Number of obs	= 734
F(3, 730)	= 205.05
Prob > F	= 0.0000
R-squared	= 0.5025
Root MSE	= .05386

coll_pooled_pooled_mean	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
par_rank_pooled_pooled_mean	.2570716	.052314	4.91	0.000	.1543679	.3597754
two_par_pooled_pooled_mean	-.2820043	.038701	-7.29	0.000	-.3579829	-.2060257
teenbrth_pooled_female_mean	-.7443224	.0694968	-10.71	0.000	-.8807598	-.607885
_cons	.564518	.0568502	9.93	0.000	.4529085	.6761275

Figure 3: Multiple linear regression results when the two control variables are included in the model. Generated by STATA.

High School Completion in the Model

To improve the model even more, a third control variable can be added. This new predictor variable is the percentage of children who completed high school or obtained a GED. This predictor is certainly relevant, as most college students who obtain a four-year degree complete high school or obtain a GED first. Therefore, it is predicted that there is a strong positive association between this predictor variable and the response variable, the percentage of children who obtain a four-year college degree. Once again, this new predictor is just a control variable. The predictor variable of interest is still the mean household income rank for parents in a commuting zone. Robust standard errors are used for this analysis, just as in previous ones.

Figure 4 below shows the results from this multivariate regression analysis with three control variables included. Every variable is considered statistically significant, even in the presence of the other regressors. Notably, the estimate for the mean household income rank for parents variable decreases to around 0.189. This suggests that if the mean household income rank for parents in a commuting zone increases by 1%, it is predicted that the percentage of children who obtain a four-year degree increases by around 0.189%, holding the other predictor variables constant. The estimate for the two parents at home predictor stays negative, which is a surprising consistency across these last two models. The estimate for a female who had a child between the ages of 13 and 19 is now about -.611. This is a slight decrease in magnitude, which suggests there might have been some slight omitted variable bias in the previous model. Finally, the estimate for percentage of children who completed high school or obtained a GED is about 0.578. This is a decently large positive association, which makes sense in a real-world sense. The traditional route for a college graduate includes graduating from high school or obtaining a GED beforehand.

Linear regression

Number of obs	734
F(4, 729)	172.56
Prob > F	0.0000
R-squared	0.5295
Root MSE	.05241

coll_pooled_pooled_mean	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
par_rank_pooled_pooled_mean	.1890244	.0525826	3.59	0.000	.085793	.2922558
two_par_pooled_pooled_mean	-.3375623	.0418671	-8.06	0.000	-.4197567	-.2553679
teenbrth_pooled_female_mean	-.6121383	.0680774	-8.99	0.000	-.7457894	-.4784873
hs_pooled_pooled_mean	.5779746	.1268086	4.56	0.000	.3290211	.8269282
_cons	.1165408	.1045222	1.11	0.265	-.0886596	.3217413

Figure 4: Multiple linear regression results when the three control variables are included in the model. Generated by STATA.

Final Multivariate Regression

For the final multivariate regression, a key control variable can be removed from the model. This removed predictor is the percentage of children claimed by two people, which is the fraction of children who had two parents at home. The previous two models claimed that there is

a negative relationship between the percentage of children having two parents at home and the percentage of children who obtained a four-year college degree. This is suspected to be false in a real world setting. In addition, there is a chance that there is some linear dependence between the two parents at home predictor and the mean household income rank for parents predictor. Generally speaking, the number of parents at home is correlated with the household's overall income. Therefore, the predictor representing the fraction of children who had two parents at home can be removed.

The finalized model has also been tested through a joint F test, which tests the model for its overall significance in explaining the response variable, the percentage of children who obtain a four year college degree. The null hypothesis for this analysis is that all of the true slope parameters for the predictor variables are all equal to zero. The alternative hypothesis is that at least one of these true slope parameters is different from zero in either direction. A standard confidence level of 5% is used for this joint F test, which contains three restrictions.

Figure 5 below shows the results of this final multivariate regression analysis and the corresponding joint F test on the model. Once again, every predictor variable is considered statistically significant, even in the presence of the other regressors. Notably, the estimate for the mean household income rank for parents variable goes back up to around 0.281. This suggests that if the mean household income rank for parents in a commuting zone increases by 1%, it is predicted that the percentage of children who obtain a four-year degree increases by around 0.281%, holding the other two predictor variables constant. For the joint F test, the F test statistic is 195.26. This is an extremely large value, and STATA reports the probability of this happening with a true null hypothesis to be 0.0000. This means that the null hypothesis is rejected. There is enough evidence from the sample data to suggest that mean household income rank for parents, percentage of females who have a child between the ages of 13 and 19, and percentage of children who graduate high school or obtain a GED all form a statistically significant model in explaining the response variable, the percentage of children who obtain a four year college degree.

Linear regression

Number of obs	734
F(3, 730)	195.26
Prob > F	0.0000
R-squared	0.4751
Root MSE	.05533

coll_pooled_pooled_mean	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
par_rank_pooled_pooled_mean	.2813068	.0528453	5.32	0.000	.17756	.3850537
teenbrth_pooled_female_mean	-.3203817	.0606549	-5.28	0.000	-.4394605	-.2013028
hs_pooled_pooled_mean	.3864729	.1214066	3.18	0.002	.1481252	.6248206
_cons	-.0791093	.1116759	-0.71	0.479	-.2983534	.1401349

Figure 5: Final multiple linear regression results when just the final two control variables are included in the model. Generated by STATA.

Nonlinear Regression

In addition to the multivariate regressions above, a nonlinear regression can be performed in an attempt to relate a predictor variable to the response variable of interest, the percentage of children who obtain a four-year college degree. The predictor chosen is the percentage of females who reported having a child between the ages of 13 and 19. The scatter plot of this variable and the dependent variable is not perfectly linear, so it is expected that there might be a quadratic relationship between them. To do this in STATA, a new variable had to be created by mathematically squaring the original female teen birth variable. From there, the original female teen birth variable and the new squared variable can be placed in a multivariate regression in STATA.

Figure 6 below shows the results of this quadratic regression in STATA. The new, quadratic variable is named 'teenbrth_squared' in STATA to avoid any confusion. Notably, both the original female teen birth variable and the new squared variable are statistically significant in the model, even in the presence of each other. The estimate for the original linear variable is -1.088355, which is certainly a strong negative correlation with the response variable, the percentage of children who obtained a four-year degree. The quadratic term has an estimate of 1.01177, which is also a large slope parameter. Notably, these two estimates have opposite signs, even though they are generated from the same variable in the data. Statistically speaking, it means that there is initially a negative correlation between the percentage of females who reported having a child as a teen and the response variable. However, at a high percentage of teenage motherhood for females, the negative effects appear to flatten out or even reverse. In a real-world sense, an area with a moderately high percentage of teenage females who have a child is predicted to have a decrease in the percentage of children who obtain a four-year college degree. However, once a high enough percentage of female teenage births is reached, the percentage of children who obtain a four-year college degree is predicted to go back up.

Source	SS	df	MS	Number of obs	=	734
				F(2, 731)	=	262.42
Model	1.77908347	2	.889541734	Prob > F	=	0.0000
Residual	2.47794838	731	.003389806	R-squared	=	0.4179
				Adj R-squared	=	0.4163
Total	4.25703185	733	.005807683	Root MSE	=	.05822

coll_pooled_pooled_mean	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
teenbrth_pooled_female_mean	-1.088355	.11866	-9.17	0.000	-1.32131	-.8553996
teenbrth_squared	1.01177	.2609634	3.88	0.000	.4994431	1.524097
_cons	.4977331	.0126196	39.44	0.000	.4729581	.5225082

Figure 6: Results of the quadratic model. Generated by STATA.

Conclusion

To conclude, every model analyzed in STATA was considered to be statistically significant in explaining the response variable, the percentage of children who obtain a four-year

degree. This means that each subset of regressor variables could be used to predict whether or not a child will have a successful outcome, which in this analysis is a four-year college degree. The most relevant subset of predictors included the average household income rank for parents, the percentage of females who have a child between the ages of 13 and 19, and the percentage of children who graduate from high school or obtain a GED. All three of these predictor variables were statistically significant in the final model, even in the presence of the other predictors. One variable, percentage of children who had two parents at home, was removed from the model over collinearity concerns. The nonlinear regression, which included a quadratic term for the percentage of females who had a child between the ages of 13 and 19, was found to have opposite effects from the linear and quadratic terms.

Like the article above mentions, it is incredibly difficult to predict an individual's outcomes just from a limited number of factors. In reality, there are most likely an infinite number of variables that go into these outcomes, and there are so many different outcomes that can be analyzed. This data analysis section looked just at the percentage of children who obtained a four-year college degree, and only four different predictor variables were analyzed for significance. This goes to show that trying to explain and predict an individual's outcomes is an extremely difficult task because of the imperfect and unlimited information that can be studied.