Intergenerational Mobility Among Graduates

The Effect of Median Household Income on Median Earnings

Empirical Economic Analysis 2 Spring 2019

Jeremiah Lam, Luan Nguyen and Rahul Thairani

Abstract

The purpose of this empirical research paper is the numerical estimation of the causal effect of median household income on median earnings of students based on data comprising averages from all degree-granting institutions in the country. The basic design of the study involves modeling several nonlinear regression specifications including control variables before performing instrumental variables regression. The findings demonstrate a low measure [0.05] for the influence of parental income on log median earnings of students, thereby implying the presence of a change in socioeconomic status between generations of the same family.[1]

[1] Williams, Yolanda. "Intergenerational Mobility: Definition & Concept." *Study.com*, Study.com, study.com/academy/lesson/intergenerational-mobility-definition-lesson-quiz.html.

Introduction

This empirical paper examines the economic issue of intergenerational mobility, which refers to the extent of dissociation between parents' and adult children's socioeconomic status, as measured by income. A strong association implies greater intergenerational transmission of advantage and therefore, less mobility.[2] On the other hand, a society with higher intergenerational mobility is one where an individual's economic status is less dependent on that of their parents.[3] Contrarily, when mobility is low, one's chances of success are primarily predetermined by birth, which can lead to unrealized human potential since talented individuals from disadvantaged families are excluded from opportunities that favor those born with privilege.[4] A higher amount of mobility is therefore critical since it provides an opportunity for children to move beyond their social origins and obtain an economic status not dictated by that of their parents.[5]

The economic question under investigation concerns what amongst graduates, is the influence of parental household income on adult student's median earnings. The null hypothesis being tested involves the absence of a causal effect of median household income on median earnings of students, implying complete intergenerational mobility, whereas the alternative hypothesis

---

[2] Fox, Liana, Florencia Torche, and Jane Waldfogel. "Intergenerational Mobility." The Oxford Handbook of the Social Science of Poverty. : Oxford University Press, April 05, 2017. Oxford Handbooks Online. Date Accessed 23 Apr. 2019 <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199914050.001.0001/oxfordhb-9780199914050-e-24>.
[3] Narayan, Ambar, and Roy Van der Weide. "Intergenerational Mobility across the World: Where Socioeconomic Status of Parents Matters the Most (and Least)." *Intergenerational Mobility across the World | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 2 July 2018, voxeu.org/article/intergenerational-mobility-across-world.
[4] Narayan, Ambar, and Roy Van der Weide. "Intergenerational Mobility across the World: Where Socioeconomic Status of Parents Matters the Most (and Least)." *Intergenerational Mobility across the World | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 2 July 2018, voxeu.org/article/intergenerational-mobility-across-world.
[5] Fox, Liana, Florencia Torche, and Jane Waldfogel. "Intergenerational Mobility." The Oxford Handbook of the Social Science of Poverty. : Oxford University Press, April 05, 2017. Oxford Handbooks Online. Date Accessed 23 Apr. 2019 <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199914050.001.0001/oxfordhb-9780199914050-e-24>.

involves its presence, and thereby estimation. The rationale for this hypothesis test concerns the 'American Dream' abstract, an ideal grounded in the United States' brand of optimism and opportunity, in which holistic freedom implies unrestricted upward social mobility.[6] This idea relies on the perception that intergenerational income mobility in the country is high since real opportunity requires mobility across generations.[7] The proposed hypothesis concerns the validation of this perception, which has come under increasing scrutiny in the last few decades. Our study concludes that the effect of median household income on median earnings of students is low, but statistically significant. This implies a relatively generous degree of intergenerational income mobility and validates the rationale for the 'American Dream' abstract. However, a significant caveat to our findings is that our dataset is only conclusive for college graduates, which has intuitive arguments towards a given ability bias or unobservable characteristics such as motivation or aptitude.

To begin with, we provide a literature review which encapsulates previous research on estimates of intergenerational mobility from various sources, followed by a description of the dataset used, before proceeding to examine this dataset for outliers and imperfect multicollinearity. After that, we model a single regressor linear specification and several nonlinear regressions including control variables before performing instrumental variables regression. Finally, we briefly summarize our findings before turning to discuss potential policy implications. The bibliography and appendices containing Stata output comprise the final components of our research paper.

---

[6] "American Dream." *Wikipedia*, Wikimedia Foundation, 9 Mar. 2019, en.wikipedia.org/wiki/American_Dream.
[7] Palomino, Juan C., et al. "Intergenerational Mobility in the US: One Size Doesn't Fit All." *Intergenerational Mobility in the US | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 3 Jan. 2019, voxeu.org/article/intergenerational-mobility-us.

Literature review

Intergenerational economic mobility is most often measured through intergenerational income elasticity in which a higher value indicates a strong association between income from generations of the same family and consequently lower mobility.[8] The first estimates of intergenerational income elasticity were around 0.2, indicating that 20% of the difference between individual income could be explained by parental income.[9] However, through the use of better databases and correcting for measurement errors, studies like Solon (1992) and Zimmerman (1992) established intergenerational income elasticity measures of around 0.4, suggesting a much higher dependence.[10] More recent studies (Palomino et al. 2018) have estimated intergenerational income mobility in the United States at approximately 0.47.

The estimated value of intergenerational income elasticity is found to decrease by 27.4% when controlling for education, which is consistent with Eide and Showalter (1999) and Cooper (2011), who find a decrease of 30% in the estimate of immobility upon its introduction.[11] Moreover, upon the addition of race in the regression specification, the estimate of intergenerational income elasticity decreases by a further 10%.[12]

---

[8] Palomino, Juan C., et al. "Intergenerational Mobility in the US: One Size Doesn't Fit All." *Intergenerational Mobility in the US | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 3 Jan. 2019, voxeu.org/article/intergenerational-mobility-us.

[9] Palomino, Juan C., et al. "Intergenerational Mobility in the US: One Size Doesn't Fit All." *Intergenerational Mobility in the US | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 3 Jan. 2019, voxeu.org/article/intergenerational-mobility-us.

[10] Palomino, Juan C., et al. "Intergenerational Mobility in the US: One Size Doesn't Fit All." *Intergenerational Mobility in the US | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 3 Jan. 2019, voxeu.org/article/intergenerational-mobility-us.

[11] Palomino, Juan C., et al. "Intergenerational Mobility in the US: One Size Doesn't Fit All." *Intergenerational Mobility in the US | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 3 Jan. 2019, voxeu.org/article/intergenerational-mobility-us.

[12] Palomino, Juan C., et al. "Intergenerational Mobility in the US: One Size Doesn't Fit All." *Intergenerational Mobility in the US | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 3 Jan. 2019, voxeu.org/article/intergenerational-mobility-us.

Data Description

Variable	Obs	Mean	Std. Dev.	Min	Max
md_earn_w~10	1,256	45119.35	11631.48	21100	124700
median_~1000	1,256	63.65795	10.71628	31.40334	95.27515
sat_avg	1,256	1060.22	136.4839	712	1555
avgfacsal	1,256	7935.963	2343.287	3052	22146
ugds_black	1,256	.1359243	.1797903	0	.9875

Table 1. Summary Statistics

The variables under consideration for any particular regression specification are:

- *md_earn_wne_p10* - median earnings of students working and not enrolled 10 years after entry - measured in dollars.
- *median_hh_inc_1000* - median household income - measured in 1000's of dollars.
- *sat_avg* - average SAT score of students admitted - measured in test score units.
- *avgfacsal* - average faculty salary - measured in dollars.
- *ugds_black* - total share of enrollment of undergraduate degree-seeking students who are black - measured in fractions.

The dataset used is Post-School Earnings Data, which comprises 2018 cross-sectional data on every degree-granting institution in the United States - trimmed to include only certain variables of interest and obtained from the College Scorecard, a public database of institution characteristics.[13] The observations with the maximum and minimum values of median earnings of students working are Albany College of Pharmacy and Health Sciences and Hebrew Theological College. The observations with the maximum and minimum values of median household income are George Mason University and Alice Lloyd College.

[13] "College Scorecard Data." *College Scorecard*, College Scorecard, 30 Oct. 2018, collegescorecard.ed.gov/data/.

Checking for Outliers

The scatterplot for the dependent variable *md_earn_wne_p10* against the key regressor *median_hh_inc_1000* has been provided in the Appendix.[14] In addition, boxplots for all variables under consideration for a particular regression specification have also been provided.[15] Despite the presence of outliers in all variables under consideration, as indicated by their boxplots, we continue to keep all observations since they pertain to the same probability distribution and therefore, represent outliers in only a statistical sense.

Checking for Imperfect Multicollinearity

The graph matrix which provides multiple pairwise scatterplots for all variables under consideration does not indicate the presence of a highly correlated relationship between any pair of regressors.[16] Additionally, the correlation table provides values for the correlation coefficients between any two variables, which range from low to moderate and thereby, indicate the absence of a close-to-linear relationship.[17] Moreover, the scatterplot between the dependent variable *md_earn_wne_p10* and the key regressor *median_hh_inc_1000* indicates the presence of a nonlinear relationship.[18] Since imperfect multicollinearity may involve several regressors, we run auxiliary regressions of one independent variable on all other independent variables, which further indicate the absence of strong collinearity since the coefficient of determination in all regression specifications lies below 0.8.[19] Lastly, all variables are individually and jointly

[14] See Figure 2 in Appendix C
[15] See Figures 3-7 in Appendix C
[16] See Figure 8 in Appendix C
[17] See Table 12 in Appendix C
[18] See Figure 2 in Appendix C
[19] See Tables 3-6 in Appendix B

statistically significant at the 5 percent significance level in a basic regression specification, as can be inferred from the respective t-tests and joint F-test, further substantiating the absence of imperfect multicollinearity.

Econometric Model

$ln\_md\_earn\_wne\_p10 = \beta_0 + \beta_1 median\_hh\_inc\_1000 + \beta_2(median\_hh\_inc\_1000)^2 +$

$\beta_3(median\_hh\_inc\_1000)^3 + \beta_4 sat\_avg + \beta_5\, ln(avgfacsal) + \beta_6 ugds\_black +$

$\beta_7(median\_hh\_inc\_1000 \times avgfacsal) + u_i$

The dependent variable in the above regression specification is *ln_md_earn_wne_p10*, the natural logarithm of the median earnings of students working, measured in dollars. The key regressor is *median_hh_inc_1000*, the median household income measured in 1000's of dollars. The polynomial terms included are $(median\_hh\_inc\_1000)^2$ and $(median\_hh\_inc\_1000)^3$, the quadratic and cubic measures of the key regressor respectively. The first control variable is *sat_avg*, the average SAT equivalent score of students admitted, measured in test score units. The second control variable is *ln(avgfacsal)*, the natural logarithm of the average faculty salary, measured in dollars. The third control variable is *ugds_black*, the total share of enrollment of undergraduate students who are black, measured in fractions. The interaction term is *(median_hh_inc_1000 × avgfacsal)*, the product of the median household income and the average faculty salary. The dependent and key independent variables are both continuous. The sign of the key coefficient is expected to be positive according to economic intuition since we expect the presence of *persistence*, defined as the intergenerational transmission of advantage. In other words, we expect greater household income to result in higher median earnings of students.

We control for other determinants of the dependent variable which are correlated with the key regressor, namely *sat_avg, ln_avgfacsal* and *ugds_black*, to reduce the omitted variable bias. The average SAT score acts as an indicator for ability, with greater ability resulting in greater median earnings of students on average. The average faculty salary serves as an indicator for institution quality, with greater quality implying increased human capital, stronger alumni networks, and improved signaling, all of which increase median earnings of students on average. The total share of enrollment of undergraduate degree-seeking students who are black acts as an indicator for racial discrimination, with a more significant fraction implying reduced job opportunities, weaker signaling and other cultural barriers that originate from within the job market.

Models of Intergenerational Mobility

Variable	model1	model2	model3	model4
median_~1000	655.893	-0.005	0.050	0.027
	24.422	0.027	0.021	0.001
sq_medi~1000		0.000	-0.001	
		0.000	0.000	
cube_me~1000		-0.000	0.000	
		0.000	0.000	
sat_avg			0.000	
			0.000	
ln_avgfacsal			0.182	
			0.045	
ugds_black			-0.147	
			0.025	
median_hh_~l			0.000	
			0.000	
_cons	3366.559	10.305	7.333	8.951
	1576.487	0.551	0.592	0.060
N	1256	1256	1256	1256
rmse	9271.293	0.176	0.137	0.227
r2	0.365	0.422	0.648	0.035
r2_a	0.365	0.421	0.646	0.035
F	721.301	305.125	328.456	855.014

legend: b/se

Table 2. Empirical Results

Single Regressor Linear Model

The first regression specification is the single regressor linear model of *md_earn_wne_p10*, the median earnings of students working and not enrolled ten years after entry regressed against the key regressor *median_hh_inc_1000*, the median household income measured in 1000's of dollars.[20]

$$md\_earn\_wne\_p10 = \beta_0 + \beta_1 median\_hh\_inc\_1000$$

The sign of the coefficient β_1 on the key regressor is positive as expected. The t-statistic of the coefficient β_1 on the key regressor is 24.42. This value of the t-statistic is higher than 1.96 in absolute value, indicating that the effect of the key regressor on the dependent variable is statistically significant at the 5 percent significance level. The economic significance of the regression lies in the magnitude of β_1 on the key regressor, which advocates that a 1000 dollar increase in parental income will result in a 655.89 dollar increase in the median earnings of students working. This result is consistent with theoretical expectations under a single regressor linear model. At a conceptual level, the critical feature justifying the hypothesis testing procedure for the coefficient β_1 on the key regressor is that, in large samples, the sampling distribution of β_1 is approximately normal according to the Central Limit Theorem. Because β_1 has a normal sampling distribution in large samples, hypotheses about the true value of the slope β_1 can be tested. Specifically, under the null hypothesis, the t-statistic is approximately distributed as a standard normal random variable. Therefore, the hypothesis can be tested at the 5 percent significance level by merely comparing the absolute value of the t-statistic to 1.96, the

[20] See Table 7 in Appendix B

critical value for a two-sided test, and rejecting the null hypothesis at the 5 percent level if the t-statistic computed is higher than 1.96. The estimated coefficients for the constant term β_0 and the slope β_1 in the single regressor linear model are 3366.56 and 655.89. The regression R-squared, defined as the fraction of the sample variance of the dependent variable predicted by all explanatory variables, is 0.365 which implies that a simple linear regression model explains 36.5% of the sample variance in the dependent variable. The standard error of the regression, defined as the estimate of the standard deviation of the regression error term, is 9271.29.

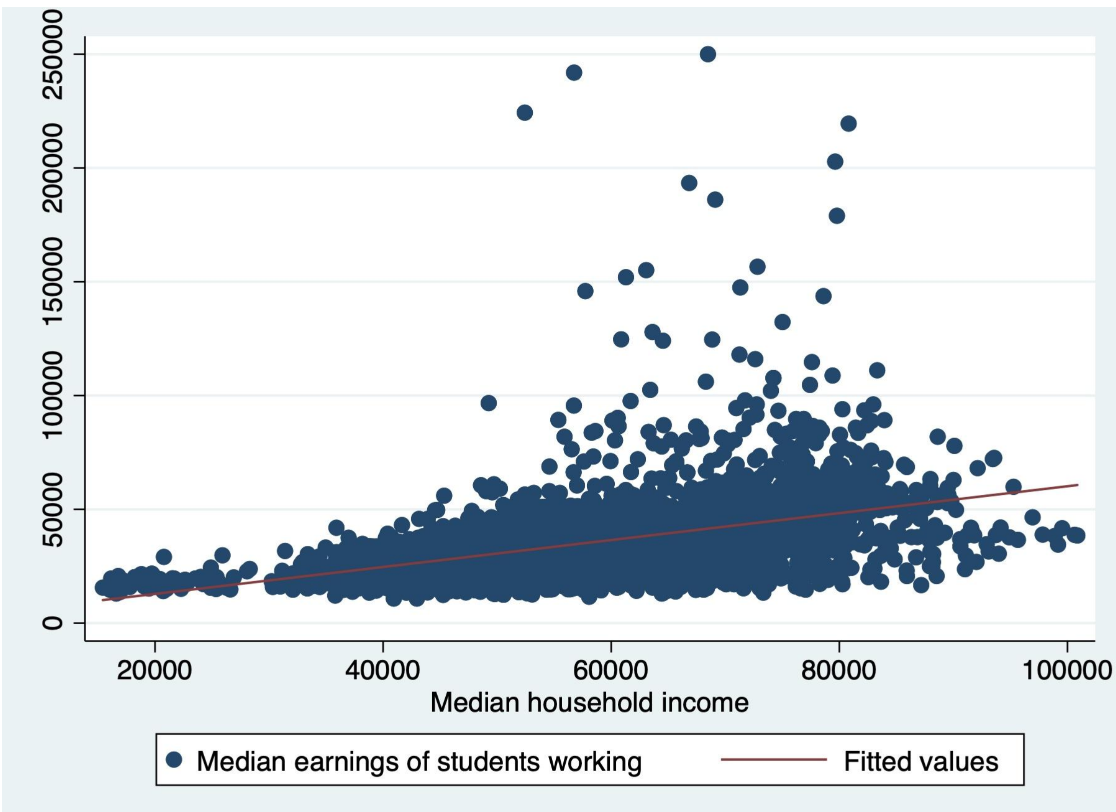


Figure 1. Scatterplot of the data with the fitted line superimposed

There appear to be nonlinearities between the dependent variable *md_earn_wne_p10* and the key regressor *median_hh_inc_1000* as evidenced by the scatterplot above. The principal nonlinearity, which takes the form of an exponential relationship between the two variables, is addressed through a log-linear model. This model better explains the variation in the dependent variable as compared to a linear model, since the natural logarithm of the dependent variable captures the positive exponential slope of X_{1i}. Additionally, the polynomial terms including the square and cubic terms X_{2i} and X_{3i} capture any additional nonlinear relationship which causes the effect of the key regressor to depend on itself.

Single Regressor Nonlinear Model

The second regression specification is the single regressor nonlinear model of *ln_md_earn_wne_p10*, the natural logarithm of the median earnings of students working regressed against the key regressor *median_hh_inc_1000*, the median household income. In addition, the regression specification contains the squared and cubic polynomial terms of the key regressor *median_hh_inc_1000*, namely $(median\_hh\_inc\_1000)^2$ and $(median\_hh\_inc\_1000)^3$ as explanatory variables.[21]

$$ln\_md\_earn\_wne\_p10 = \beta_0 + \beta_1 median\_hh\_inc\_1000 + \beta_2(median\_hh\_inc\_1000)^2 + \beta_3(median\_hh\_inc\_1000)^3.$$

The t-statistic of the coefficient β_1 on the key regressor is -2.40. This value of the t-statistic is greater than 1.96 in absolute value, indicating that the effect of the key regressor on the dependent variable is still statistically significant at the 5 percent significance level. The regression R-squared is 0.422; however, this result is incomparable to the initial regression since the dependent variable in both regression specifications is different. The standard error of

[21] See Table 8 in Appendix B

regression for the single regressor nonlinear model is 0.176. There are several determinants of the dependent variable *md_earn_wne_p10* which are correlated with the key regressor *median_hh_inc_1000*, resulting in omitted variable bias. To mitigate this bias, we control for several of these determinants.

Multiple Regressor Nonlinear Model

The third regression specification is the multiple regressor nonlinear model of *ln_md_earn_wne_p10,* the natural logarithm of the median earnings of students working regressed against the key regressor of *median_hh_inc_1000*, the median household income. In addition, the regression specification contains the squared and cubic polynomial terms of the key regressor and additional control variables in the form of *sat_avg*, the average SAT score of students admitted, *ln(avgfacsal)*, the natural logarithm of the average faculty salary and *ugds_black*, the total share of enrollment of undergraduate degree-seeking students who are black. Lastly, we also include the interaction term (*median_hh_inc_1000 × avgfacsal)*, the product of the median household income and the average faculty salary, which allows for the effect of the parental income on median earnings of students working and not enrolled to depend on the average faculty salary.[22]

$$ln\_md\_earn\_wne\_p10 = \beta_0 + \beta_1 median\_hh\_inc\_1000 + \beta_2(median\_hh\_inc\_1000)^2 + \beta_3(median\_hh\_inc\_1000)^3 + \beta_4 sat\_avg + \beta_5\, ln(avgfacsal) + \beta_6 ugds\_black + \beta_7(median\_hh\_inc\_1000 \times avgfacsal)$$

[22] See Table 9 in Appendix B

The sign of the coefficient β_1 on the key regressor is positive. The t-statistic of the coefficient β_1 on the key regressor is 2.36. This value of the t-statistic is greater than 1.96 in absolute value, indicating that the effect of the key regressor on the dependent variable is statistically significant at the 5 percent significance level. In the multiple regressor nonlinear model, the coefficient β_1 of the key regressor does not have a natural interpretation as in the multiple regressor linear model. For instance, it is not very helpful to think of β_1 as the effect of changing the median household income, keeping the square of the median household income constant. The magnitude of the slope coefficient β_1 changes significantly from -0.005 in the single regressor nonlinear model to 0.050 in the multiple regressor nonlinear model, which indicates the presence of omitted variable bias in the single regressor nonlinear model. The estimates for the coefficients are 0.050 for β_1 on the key regressor of *median_hh_inc_1000*, -0.001 for β_2 on the regressor *(median_hh_inc_1000)*^{2}, 3.48e-06 for β_3 on the regressor *(median_hh_inc_1000)*^{3} and 2.34e-07 for β_7 on the interaction term *(median_hh_inc_1000 × avgfacsal)*. The polynomial and interaction terms including the key regressor have t-statistics that are greater than 1.96 in absolute value, which indicates that the effect of all such terms on the dependent variable is statistically significant at the 5 percent significance level. The coefficient β_2 on the squared polynomial term *(median_hh_inc_1000)*^{2} and the coefficient β_3 on the cubic polynomial term *(median_hh_inc_1000)*^{3} do not have a natural interpretation as in the multiple regressor linear model. Therefore, the coefficients on the squared and cubic polynomial terms do not have economic significance. The coefficient β_7 on the interaction term *(median_hh_inc_1000 × avgfacsal)* is interpreted as the increase in the magnitude of the causal effect of median household income on median earnings of students for a unit increase in average faculty salary.

To estimate the polynomial regression model for a single regressor using the sequential hypothesis testing procedure, we began by including polynomial terms up until the cubic form of the key regressor. Since the coefficient on the cubic term was statistically significant and the coefficient on the key regressor of interest was positive, we incorporated the cubic term in the regression specification, thereby concluding the sequential hypothesis testing procedure. From our model, we conclude with 95% confidence that the true population parameter for the effect of median household income measured in 1000's of dollars on the natural logarithm of the median earnings of students lies between 0.0083 and 0.0912. The advantage of reporting the key effect as an interval estimate instead of a point estimate is that interval estimation can provide a range of values with a known probability of capturing the population parameter of the key effect. Therefore, interval estimation can lend a proper perspective concerning the interpretation of the key effect.

The coefficients of the control variables *sat_avg*, *ln(avgfacsal)* and *ugds_black* meet expectations about their sign. The coefficients of *sat_avg* and *ln(avgfacsal)* are positive whereas the coefficient on *ugds_black* is negative. The coefficients of the control variables do not have a causal interpretation since their inclusion is primarily justified on the grounds of mitigating omitted variable bias in the estimate of the coefficient on the key regressor. The coefficients on the control variables all have t-statistics that are greater than 1.96 in absolute value, indicating that all control variables are statistically significant at the 5 percent significance level. The regression R-squared increases from 0.422 in the single regressor nonlinear model to 0.6482 in the multiple regressor nonlinear model while the adjusted R-squared increases from 0.421 to 0.6462 respectively. The standard error of regression decreases from 0.176 in the single regressor nonlinear model to 0.137 in the multiple regressor nonlinear model. The above-mentioned changes in the goodness-of-fit measures indicate an improvement in the fit of the model. The result of the F-test for the joint significance of slope coefficients is 328.46. Comparing this value to $F_{7,\infty}$, we can conclude that the coefficients on all regressors included in the multiple regressor nonlinear specification are jointly statistically significant at the 5 percent significance level. To test the linearity of this regression specification is linear, we test the following null and alternative hypotheses.[23]

$H_0 : \beta_2 = 0, \beta_3 = 0$ *vs.* $H_1 : \beta_j \neq 0$, *at least one* j, $j=1, 2$

The F-statistic for the above joint hypothesis test is 2.68. Comparing this to the value of $F_{2,\infty}$, we can conclude that the coefficients on the squared and cubic terms of the polynomial regression specification are jointly statistically significant at the 10 percent significance level.

In spite of the inclusion of control variables in the multiple regressor nonlinear specification, there continues to exist omitted variable bias in the coefficient β_1 on the key regressor, since data on all variables that are determinants of the dependent variable *ln_md_earn_wne_p10*, the natural logarithm of the median earnings of students working and correlated with the key regressor *median_hh_inc_1000*, median household income, is not available. In order to further mitigate this omitted variable bias, arising from either omitted variables, errors in variables or

[23] See Table 13 in Appendix D

simultaneous causality bias, when causality runs both from X_{1i} to Y_i and from Y_i to X_{1i}, we turn to instrumental variables regression.

Instrumental Variables Regression Model

The fourth regression specification is the instrumental variables regression of *ln_md_earn_wne_p10*, the natural logarithm of the median earnings of students working and not enrolled regressed against *median_hh_inc_1000*, the median household income measured in 1000's of dollars. The instrumental variables included in this regression specification are *sat_avg*, the average SAT equivalent score of students admitted, *ln(avgfacsal)*, the natural logarithm of the average faculty salary and *ugds_black*, the total share of enrollment of undergraduate degree-seeking students who are black.[24] The sign of the coefficient β_1 on the key regressor is positive. The t-statistic of the coefficient β_1 on the key regressor is 29.24. This value of the t-statistic is greater than 1.96 in absolute value, indicating that the effect of the key regressor on the dependent variable is statistically significant at the 5 percent significance level. In this instrumental variables regression specification, which takes the form of a log-linear model, the coefficient β_1 on the key regressor of 0.027 implies that a 1000 dollar increase in the median household income results in a 2.7% increase in the median earnings of students working and not enrolled. The coefficient β_1 on the key regressor changes from 0.05 in the multiple regressor nonlinear model to 0.027 in the instrumental variables regression model. In order to test the validation of the instrument relevancy condition, we run several regressions of the key regressor on a particular instrumental variable. The t-statistic of the coefficient β_1 on the key

[24] See Tables 10-11 in Appendix B

regressor in all such specifications is greater than 1.96 in absolute value, indicating that the instruments are statistically significant determinants of the key regressor at the 5 percent significance level. Finally, the instrumental variables regression model does not include any exogenous regressors.

The elasticity of the dependent variable *md_earn_wne_p10*, the median earnings of students working and not enrolled with respect to the key regressor *median_hh_inc*, the median household income is 0.925.[25] The computed elasticity measure implies that a 1 percent increase in median household income results in a 0.925 percent increase in the median earnings of students working and not enrolled. This unitary elastic measure is not small enough for us to disregard in an economic sense, the relationship between the dependent variable and the key regressor.

Summary and Potential Extensions

This empirical research paper models four different regression specifications in order to estimate a numerical measure for the causal effect of the key regressor *median_hh_inc_1000* on the dependent variable *md_earn_wne_p10*. The coefficient β_1 on the key regressor changes in magnitude with every subsequent model, indicating the presence of omitted variable bias. The multiple regressors nonlinear model yields the most precise estimate for coefficient β_1 on the key regressor, given the scope of this dataset. Specifically, the estimated value of the coefficient β_1 on the key regresor in the multiple regressors nonlinear model is 0.050. The coefficient β_1 on the key regressor in all regression specifications has a t-statistic greater than 1.96 in absolute value,

[25] See Table 14 in Appendix D

which indicates that all such coefficients are statistically significant at the 5 percent significance level. These results are consistent with the initial expectation that parental income has a positive influence on the median earnings of students. An important limitation of the dataset being used is that it only contains data on college graduates in the United States. The extent of this dataset therefore, limits the ability of any particular regression specification to estimate a coefficient consistent with those determined by past studies. Nevertheless, these regression models are still applicable in that they estimate a statistically significant and positive relationship between the key regressor and the dependent variable.

Potential future analyses regarding this topic can begin by selecting a more comprehensive dataset that includes population groups not examined by our empirical research. This inclusion would allow for a more accurate estimate of the coefficient β_1 on the key regressor. Moreover, the regression models used in this empirical paper are rudimentary when compared with other advanced statistical models. Future investigations can apply such models, thereby accounting for additional relationships and correcting for measurement errors. The connection between median household income and median earnings of students ties into the idea of a 'poverty trap', where some families do not have the amount of income required for their children to be afforded the opportunities needed in order to move up the socioeconomic ladder. By boosting initial assets either through subsidy or other forms of aid an important source of unequal opportunities – namely, parental income transmission – will be curbed. The education system is one of the key channels for the transmission of income across generations, and therefore is one of the main sectors in which such tailored policies should be implemented.

Bibliography

"American Dream." *Wikipedia*, Wikimedia Foundation, 9 Mar. 2019, en.wikipedia.org/wiki/American_Dream.

Cooper, D P (2011), "Unlocking the American dream: exploring intergenerational social mobility and the persistence of economic status in the United States".

Eide, E R and M H Showalter (1999), "Factors affecting the transmission of earnings across generations: A quantile regression approach", *Journal of Human Resources* 34(2): 253–267.

Fox, Liana, Florencia Torche, and Jane Waldfogel. "Intergenerational Mobility." The Oxford Handbook of the Social Science of Poverty. : Oxford University Press, April 05, 2017. Oxford Handbooks Online. Date Accessed 23 Apr. 2019

Narayan, Ambar, and Roy Van der Weide. "Intergenerational Mobility across the World: Where Socioeconomic Status of Parents Matters the Most (and Least)." *Intergenerational Mobility across the World | VOX, CEPR Policy Portal*, VOX, CEPR Policy Portal, 2 July, 2018.

Palomino, J C, G Marrero and J G Rodríguez (2018), "One size doesn't fit all: a quantile analysis of intergenerational income mobility in the U.S. (1980-2010)", *Journal of Economic Inequality* 16(3): 347-367.

Solon, G (1992), "Intergenerational income mobility in the United States", *The American Economic Review* 82(3): 393–408.

Yolanda, William. "Intergenerational Mobility: Definition & Concept." *Study.com*, Study.com, 2018, study.com/academy/lesson/intergenerational-mobility-definition-lesson-quiz.html.

Zimmerman, D J (1992), "Regression toward mediocrity in economic stature", *The American Economic Review* 82(3): 409–429.

Appendix A

<u>Stata do-file</u>

```
use projectdata.dta

// obtain the universal sample for all regression specifications
quietly regress ln_md_earn_wne_p10 median_hh_inc_1000 sq_median_hh_inc_1000
cube_median_hh_inc_1000 sat_avg ln_avgfacsal ugds_black median_hh_inc_1000_avgfacsal

// summarize all variables under universal sample
summarize md_earn_wne_p10 median_hh_inc_1000 sat_avg avgfacsal ugds_black if e(sample)
== 1

// scatterplot of dependent variable md_earn_wne_p10 against key regressor
median_hh_inc_1000 along with linear regression line fitted
twoway scatter md_earn_wne_p10 median_hh_inc_1000 || lfit md_earn_wne_p10
median_hh_inc_1000

// boxplots for all variables under consideration for any particular regression specification
graph box md_earn_wne_p10
graph box median_hh_inc_1000
graph box sat_avg
graph box avgfacsal
graph box ugds_black

// graph matrix for all variables under consideration for any particular regression specification
graph matrix md_earn_wne_p10 median_hh_inc_1000 sat_avg avgfacsal ugds_black

// correlation table for all variables under consideration for any particular regression specification
corr md_earn_wne_p10 median_hh_inc sat_avg avgfacsal ugds_black

// auxiliary regressions of one independent variable against all other independent variables
regress median_hh_inc_1000 sat_avg avgfacsal ugds_black
regress sat_avg median_hh_inc_1000 avgfacsal ugds_black
regress avgfacsal median_hh_inc_1000 sat_avg ugds_black
regress ugds_black sat_avg median_hh_inc_1000 avgfacsal

// creation of empirical results output table
```

quietly regress md_earn_wne_p10 median_hh_inc_1000 if e(sample) == 1
estimates store model1
quietly regress ln_md_earn_wne_p10 median_hh_inc_1000 sq_median_hh_inc_1000 cube_median_hh_inc_1000 if e(sample) == 1
estimates store model2
quietly regress ln_md_earn_wne_p10 median_hh_inc_1000 sq_median_hh_inc_1000 cube_median_hh_inc_1000 sat_avg ln_avgfacsal ugds_black median_hh_inc_1000_avgfacsal if e(sample) == 1
estimates store model3
quietly ivreg ln_md_earn_wne_p10 (median_hh_inc_1000 = ln_avgfacsal ugds_black sat_avg), r
estimates store model4
estimates table model1 model2 model3 model4, b(%9.3f) se(%6.3f) stats(N rmse r2 r2_a F) title("Models of Intergenerational Mobility")

// single regressor linear model
regress md_earn_wne_p10 median_hh_inc_1000 if e(sample) ==1

// single regressor nonlinear model
regress ln_md_earn_wne_p10 median_hh_inc_1000 sq_median_hh_inc_1000 cube_median_hh_inc_1000 if e(sample) ==1

// multiple regressor nonlinear model
regress ln_md_earn_wne_p10 median_hh_inc_1000 sq_median_hh_inc_1000 cube_median_hh_inc_1000 sat_avg ln_avgfacsal ugds_black median_hh_inc_1000_avgfacsal if e(sample) ==1

// F-test for linear vs. nonlinear model
test sq_median_hh_inc_1000 cube_median_hh_inc_1000

// instrumental variables model
ivreg ln_md_earn_wne_p10 (median_hh_inc_1000 = ln_avgfacsal ugds_black sat_avg), r

// computation of elasticity of dependent variable with respect to key regressor
quietly regress md_earn_wne_p10 median_hh_inc
margins if e(sample) ==1, eyex(median_hh_inc) atmeans

Appendix B

Table 3. Auxiliary Regression 1

Source	SS	df	MS			
Model	53859.3986	3	17953.1329	Number of obs	=	1,270
Residual	98299.9262	1,266	77.6460712	F(3, 1266)	=	231.22
				Prob > F	=	0.0000
				R-squared	=	0.3540
				Adj R-squared	=	0.3524
Total	152159.325	1,269	119.904905	Root MSE	=	8.8117

median_~1000	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat_avg	.0165854	.0028019	5.92	0.000	.0110885	.0220823
avgfacsal	.0018225	.0001493	12.21	0.000	.0015296	.0021155
ugds_black	-5.765896	1.568257	-3.68	0.000	-8.842564	-2.689228
_cons	32.36628	2.441859	13.25	0.000	27.57574	37.15681

Table 4. Auxiliary Regression 2

Source	SS	df	MS			
Model	14154210	3	4718070	Number of obs	=	1,270
Residual	9623926.69	1,266	7601.83783	F(3, 1266)	=	620.65
				Prob > F	=	0.0000
				R-squared	=	0.5953
				Adj R-squared	=	0.5943
Total	23778136.7	1,269	18737.6964	Root MSE	=	87.189

sat_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
median_hh_inc_1000	1.623769	.2743182	5.92	0.000	1.085601	2.161937
avgfacsal	.0327226	.0012626	25.92	0.000	.0302457	.0351996
ugds_black	-221.0657	14.30931	-15.45	0.000	-249.1382	-192.9931
_cons	727.8462	15.69485	46.37	0.000	697.0554	758.637

Table 5. Auxiliary Regression 3

Source	SS	df	MS
Model	3.8874e+09	3	1.2958e+09
Residual	3.1157e+09	1,266	2461071.79
Total	7.0032e+09	1,269	5518642.66

Number of obs	=	1,270
F(3, 1266)	=	526.52
Prob > F	=	0.0000
R-squared	=	0.5551
Adj R-squared	=	0.5540
Root MSE	=	1568.8

avgfacsal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
median_hh_inc_1000	57.76688	4.732915	12.21	0.000	48.48166	67.0521
sat_avg	10.59386	.4087483	25.92	0.000	9.791957	11.39575
ugds_black	1611.311	277.012	5.82	0.000	1067.858	2154.764
_cons	-7203.897	417.4067	-17.26	0.000	-8022.782	-6385.012

Table 6. Auxiliary Regression 4

Source	SS	df	MS
Model	9.48253449	3	3.16084483
Residual	31.2372416	1,266	.024673967
Total	40.7197761	1,269	.032088082

Number of obs	=	1,270
F(3, 1266)	=	128.10
Prob > F	=	0.0000
R-squared	=	0.2329
Adj R-squared	=	0.2311
Root MSE	=	.15708

ugds_black	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat_avg	-.0007175	.0000464	-15.45	0.000	-.0008087	-.0006264
median_hh_inc_1000	-.0018323	.0004984	-3.68	0.000	-.0028099	-.0008546
avgfacsal	.0000162	2.78e-06	5.82	0.000	.0000107	.0000216
_cons	.8845373	.0392393	22.54	0.000	.8075561	.9615185

Table 7. Single Regressor Linear Model

Source	SS	df	MS			
Model	6.2001e+10	1	6.2001e+10	Number of obs	=	1,256
Residual	1.0779e+11	1,254	85956882.6	F(1, 1254)	=	721.30
				Prob > F	=	0.0000
Total	1.6979e+11	1,255	135291426	R-squared	=	0.3652
				Adj R-squared	=	0.3647
				Root MSE	=	9271.3

md_earn_wne_p10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
median_hh_inc_1000	655.8928	24.42162	26.86	0.000	607.9811	703.8045
_cons	3366.559	1576.487	2.14	0.033	273.715	6459.402

Table 8. Single Regressor Nonlinear Model

Source	SS	df	MS			
Model	28.2709122	3	9.4236374	Number of obs	=	1,256
Residual	38.6673866	1,252	.030884494	F(3, 1252)	=	305.13
				Prob > F	=	0.0000
Total	66.9382988	1,255	.05333729	R-squared	=	0.4223
				Adj R-squared	=	0.4210
				Root MSE	=	.17574

ln_md_earn_wne_p10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
median_hh_inc_1000	-.0050195	.0268914	-0.19	0.852	-.0577766	.0477376
sq_median_hh_inc_1000	.0002126	.0004296	0.49	0.621	-.0006302	.0010555
cube_median_hh_inc_1000	-6.54e-07	2.25e-06	-0.29	0.771	-5.06e-06	3.76e-06
_cons	10.30534	.5511133	18.70	0.000	9.224132	11.38655

Table 9. Multiple Regressor Nonlinear Model

Source	SS	df	MS
Model	43.387571	7	6.19822442
Residual	23.5507279	1,248	.018870776
Total	66.9382988	1,255	.05333729

Number of obs	=	1,256
F(7, 1248)	=	328.46
Prob > F	=	0.0000
R-squared	=	0.6482
Adj R-squared	=	0.6462
Root MSE	=	.13737

ln_md_earn_wne_p10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
median_hh_inc_1000	.0497822	.021136	2.36	0.019	.0083162	.0912483
sq_median_hh_inc_1000	-.0006987	.0003378	-2.07	0.039	-.0013614	-.000036
cube_median_hh_inc_1000	3.48e-06	1.77e-06	1.97	0.049	1.40e-08	6.95e-06
sat_avg	.0003737	.0000458	8.16	0.000	.0002838	.0004636
ln_avgfacsal	.18191	.0453827	4.01	0.000	.0928751	.2709449
ugds_black	-.1465463	.0250183	-5.86	0.000	-.1956289	-.0974638
median_hh_inc_1000_avgfacsal	2.34e-07	8.18e-08	2.86	0.004	7.38e-08	3.95e-07
_cons	7.332535	.5920534	12.38	0.000	6.171005	8.494065

Table 10. Instrumental Variables First-Stage Regression

First-stage regressions

Source	SS	df	MS
Model	53243.9964	3	17747.9988
Residual	90878.4494	1,252	72.5866209
Total	144122.446	1,255	114.838602

Number of obs	=	1,256
F(3, 1252)	=	244.51
Prob > F	=	0.0000
R-squared	=	0.3694
Adj R-squared	=	0.3679
Root MSE	=	8.5198

median_~1000	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat_avg	.0187979	.0026092	7.20	0.000	.0136789	.0239169
ln_avgfacsal	14.69513	1.178633	12.47	0.000	12.38281	17.00744
ugds_black	-5.343764	1.511965	-3.53	0.000	-8.310028	-2.3775
_cons	-86.91658	8.999025	-9.66	0.000	-104.5714	-69.26175

Table 11. Instrumental Variables Second-Stage Regression

```
Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =     1,256
-------------+----------------------------------   F(1, 1254)      =    770.99
       Model |  2.36711492         1  2.36711492   Prob > F        =    0.0000
    Residual |  64.5711839     1,254  .051492172   R-squared       =    0.0354
-------------+----------------------------------   Adj R-squared   =    0.0346
       Total |  66.9382988     1,255   .05333729   Root MSE        =    .22692

------------------------------------------------------------------------------------
ln_md_earn_wne_p10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
median_hh_inc_1000 |   .0273061   .0009834    27.77   0.000     .0253768    .0292354
             _cons |   8.950616   .0629285   142.23   0.000     8.827159    9.074073
------------------------------------------------------------------------------------
Instrumented:  median_hh_inc_1000
Instruments:   sat_avg ln_avgfacsal ugds_black
------------------------------------------------------------------------------------
```

Appendix C

Figure 2. Scatterplot of the Data with the Fitted Line Superimposed

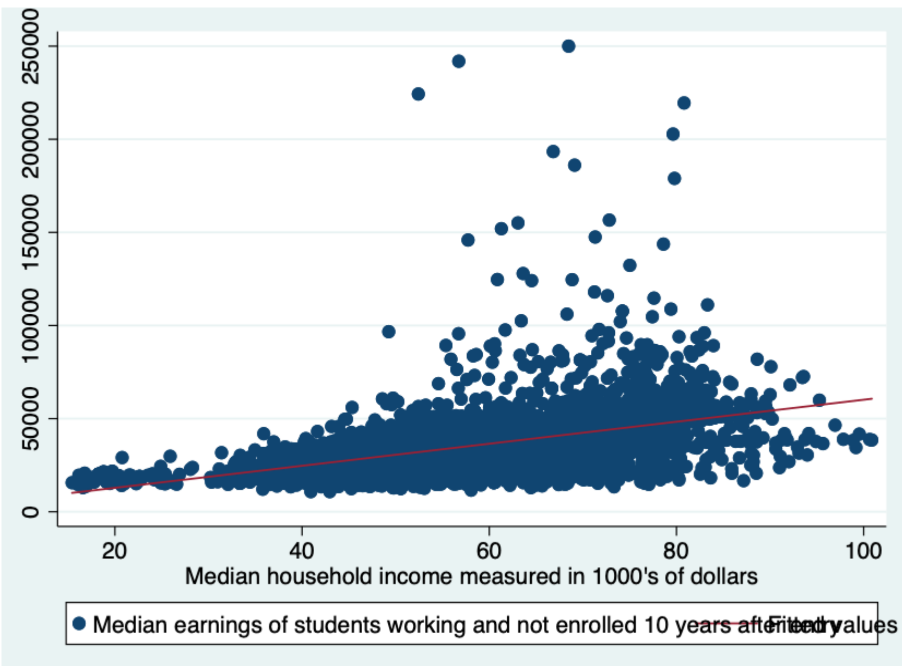


Figure 3. Boxplot of Median Earnings of Students Working and Not Enrolled

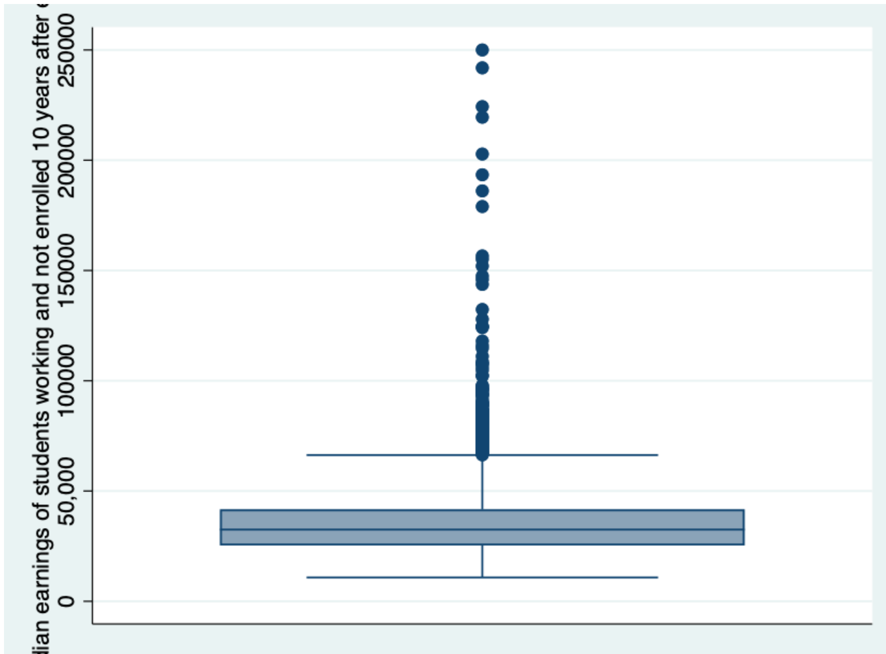


Figure 4. Boxplot of Median Household Income Measured in 1000’s of Dollars

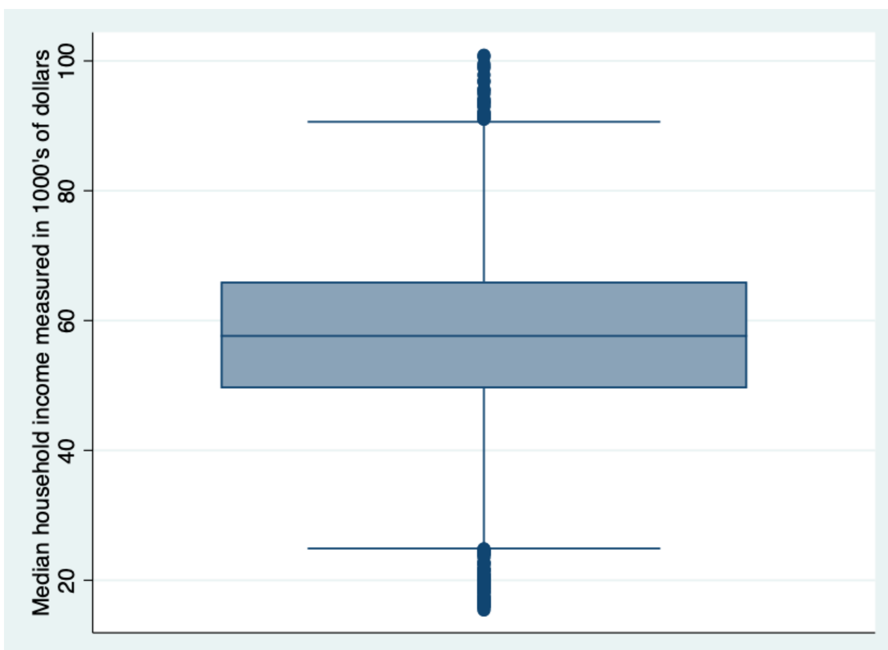


Figure 5. Boxplot of Average SAT Equivalent Score of Students Admitted

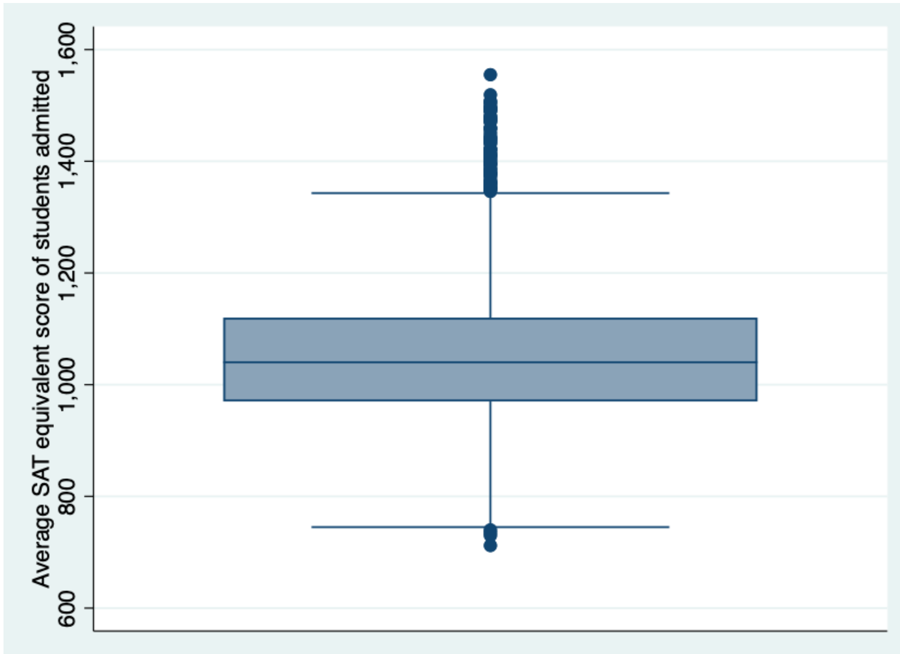


Figure 6. Boxplot of Average Faculty Salary

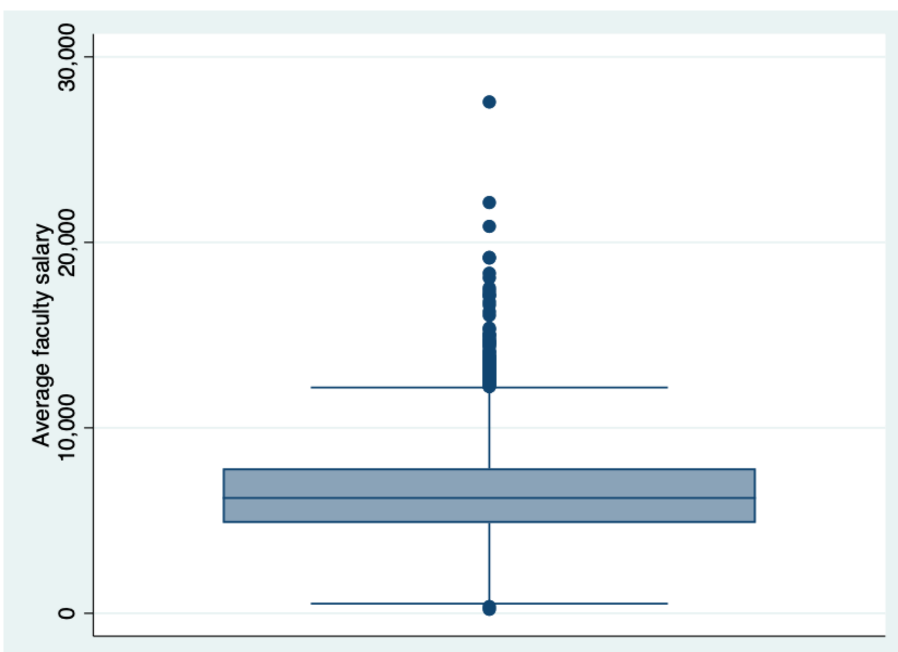


Figure 7. Boxplot of Total Share of Undergraduate Degree-Seeking Students who are Black

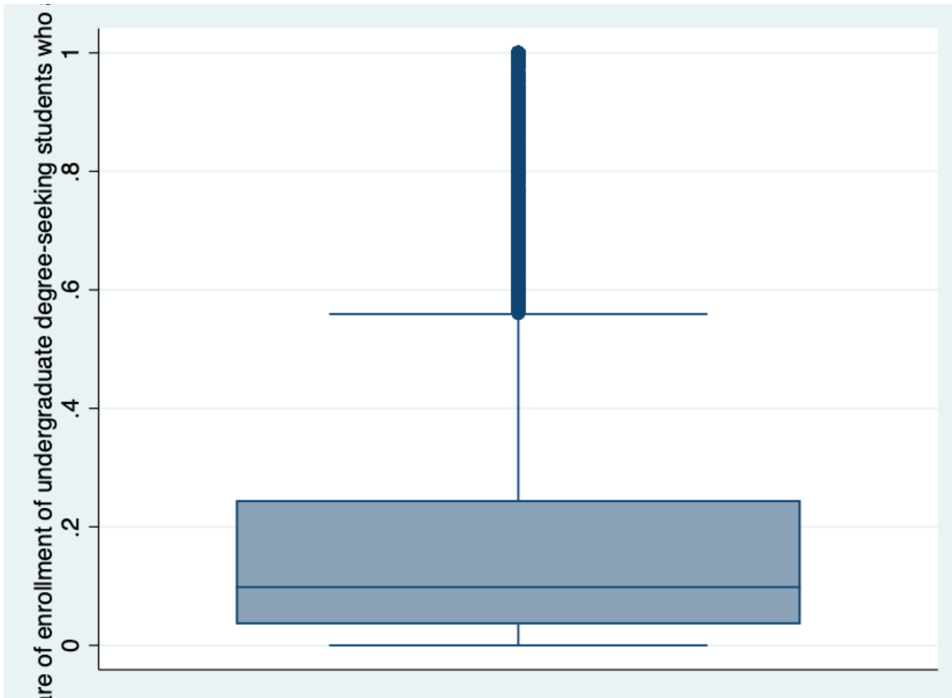
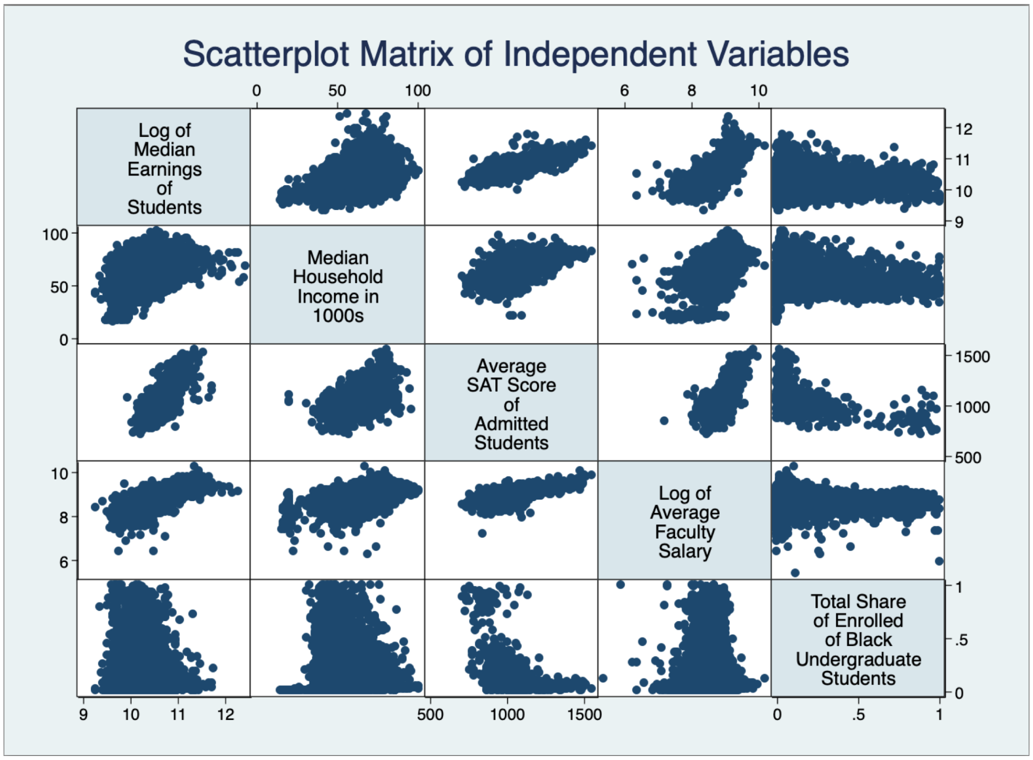


Table 12. Correlation Table

	med~1000	sat_avg	ln_avg~l	ugds_b~k
median_~1000	**1.0000**			
sat_avg	**0.5368**	**1.0000**		
ln_avgfacsal	**0.5641**	**0.6739**	**1.0000**	
ugds_black	**-0.2951**	**-0.4594**	**-0.2511**	**1.0000**

Figure 8. Graph Matrix



Appendix D

Table 13. Joint Hypothesis Testing Linear vs. Nonlinear Model

```
 ( 1)  sq_median_hh_inc_1000 = 0
 ( 2)  cube_median_hh_inc_1000 = 0

       F(  2,  1248) =    2.68
            Prob > F =    0.0687
```

Table 14. Elasticity of Median Earnings of Students with respect to Median Household Income

```
Conditional marginal effects                    Number of obs     =      1,256
Model VCE    : OLS

Expression   : Linear prediction, predict()
ey/ex w.r.t. : median_hh_inc
at           : median_hh_~c    =    63657.95 (mean)
```

	ey/ex	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
median_hh_inc	.9253855	.0348712	26.54	0.000	.8569731	.9937978