An Investigation into the Relationship between Household Socio-Economic Factors and Children's ACT Score

Kristof Pusztai
Department of Statistics
Columbia University
1255 Amsterdam Ave, New York, NY 10027
k.pusztai@columbia.edu

December 2021

1 Introduction

There are a wide variety of studies which show that education and financial success are highly linked[1][2]. These studies focus on looking at the correlation between educational attainment level and income and have independently found high correlations. Whether this correlation can be labelled as a causal relationship is still up for debate as there are many other factors at play which determine income levels. However, in today's world where education plays a huge part in the types of jobs one can attain, it is unquestionable that there is a significant relationship between the two. This naturally leads us to another question, what sorts of factors actually contribute to higher educational attainment?

It has been argued that "education attainment is a continuous process in which the education achievement of the prior stage affects the later-stage achievement both cumulatively and probabilistically"[3]. In fact, our entire education system is built on trying to quantify education achievement and performance via standardized measures such as GPA and standardized test scores. These measurements significantly affect the ability of the student to pursue further, higher level education. As a result, investigating factors which affect these educational "success" scores inherently gives us insight into the factors which contribute to educational attainment later down the line.

In this paper, we examine the relationship between certain household factors and ACT scores which are a metric widely used in college admissions. Some initial exploratory data analysis shows us that the distribution of ACT Composite scores among the different states follows a somewhat normal distribution. This can be seen in the Q-Q plot in Figure 1 which does exhibit signs of short-

tailedness. However, this is expected as we have upper and lower bounds on the ACT score so our tails can only extend so far. To make our normality assumption more formal, we perform a Kolmogorov-Smirnov test on our composite score data with a null hypothesis that our data comes from a Normal distribution. We find a p-value of 0.3237 and, as a result, fail to reject the null hypothesis at a significance level of $\alpha = 0.05$. Additionally, we observe some interesting yet not surprising trends in the correlation matrix found in Figure 2. Specifically, we note that the factors bachelors or higher, different house, public transport, carpooled, mean travel time, employment industry, and income levels all have a strong correlation with the ACT composite score. Some interesting inter-factor correlations exist as well such as a strong negative correlation between higher education and living in a different house one year ago. We also see some strong correlation between median wage and higher education which agrees with the studies mentioned earlier. We will keep these inter-factor correlations in mind as this multi-collinearity has potential to affect our inferential results.

2 Data Collection and Data Description

All data was downloaded and merged from the National Center for Education Statistics database. The downloads of individual excel files were done manually and a a python script (data_merge.ipynb) was used to merge all the data into a final csv file. Our ACT composite score data was downloaded from a table which contains the mean for each individual state in 2019[4]. Our household factor data was downloaded individually for each of the 50 states plus Washington as the NCES database did not have a cumulative data file[5]. The population of interest used in the NCES data retrieval query was "Parents of Relevant Children – Enrolled" with the "Social" and "Economic" tables containing the relevant co-variate data.

After merging all of our collected data, we are left with 51 samples each with 21 variables. Narrowing these variables down, we are interested in general socio-economic household characteristics such as education level of parents, residence in the previous year (moving around often implies more instability), marriage status, language(s) spoken at home, and household income. For count data such as marriage status, or education higher then bachelors, we use percentage instead of the raw count data. This is important because our model parameters will be related to change in percentage instead of absolute count which we believe is more intuitive and useful to gain an understanding of any existing relationships. For numerical data such as household income, we use non-segmented[1] median income as our co-variate as this reflects the earnings of both men and women in our population of interest.

[1]by non-segmented, we mean the median calculation includes data from both employed men and women

3 Statistical Model

Since we are interested in inference on certain variables rather than prediction, no extensive model selection was performed. Instead, we performed several regressions which test the significance of our covariates both individually and together. Our main model of interest is the full model which contains all covariates together, however, from the corellation matrix in Figure 2 we note that some variables will likely suffer from the effects of collinearity. As a result, to examine individual significance we also performed individual regressions. Our full model along with the R output converted to LaTeX table format(parsed and converted via texreg library) can be seen in Table 1. We had a total of 6 models with functional forms as follows[2]:

$$y = \texttt{average composite score}$$

$$x_1 = \texttt{count of bachelors or higher}$$

$$x_2 = \texttt{count of living in a different house 1 year ago}$$

$$x_3 = \texttt{count of married}$$

$$x_4 = \texttt{count of other language at home}$$

$$x_5 = \texttt{median wage in units of \$10,000}$$

Model 1:

$$y = \beta_0 + \beta_1 * x_1$$

Model 2:

$$y = \beta_0 + \beta_1 * x_2$$

Model 3:

$$y = \beta_0 + \beta_1 * x_3$$

Model 4:

$$y = \beta_0 + \beta_1 * x_4$$

Model 5:

$$y = \beta_0 + \beta_1 * x_5$$

Model 6 (Full):

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + \beta_5 * x_5$$

[2]counts are for relevant population as specified in NCES query

	Estimated Coefficients (Std. Error)
(Intercept)	16.88
	(9.79)
bachelors_higher	0.16^{*}
	(0.06)
diff_house	-0.51^{***}
	(0.13)
married	0.04
	(0.11)
other_language	0.02
	(0.02)
median_wage_scaled	0.03
	(0.60)
R^2	0.67
Adj. R^2	0.64
Num. obs.	51

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Full Model

4 Research Question

From Table 1-6, we see the results of the T-tests performed by R. The numbers in the parenthesis denote the standard error of the estimated coefficients. From the full model regression output in Table 1, we see that only two of the estimated coefficients are deemed significant at an $\alpha = 0.05$ significance level. The two significant coefficients correspond to the count of bachelors or higher and the count of those who lived in a different house 1 year ago. This implies that having a bachelors or higher has a statistically significant positive relationship with ACT score, and living in a different house 1 year ago has a negative relationship. The rest of our covariates in the full model are not significant via the t-test at significance of $\alpha = 0.05$.

However, we must be careful here to discard these other variables as from our individual regression t-tests, we see that all variables except speaking another language at home are deemed significant at the $\alpha = 0.05$ level. As a result, we cannot conclude that marriage status and median wage do not have a statistically significant relationship with respect to ACT score. In fact, we note from our earlier exploratory data analysis that there are strong positive correlations between bachelors or higher and both marriage and median wage covariates. The negative effects of this co-linearity are reflected in the standard errors which result in the non-significance of these variables in our full model. The co-linearity has lowered the power of our model and the relationships between ACT score and marriage and median wage should not be dismissed.

5 Appendix

5.1 Diagnostics and Model Validation

5.1.1 Diagnostic Plots

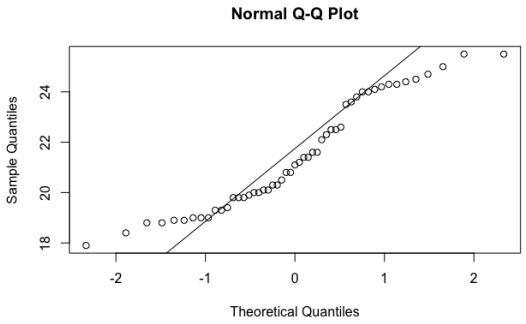


Figure 1: QQ-Plot of ACT Composite score data. We see that besides some deviation at the upper and lower ends, visually, our data seems to satisfy our normality assumption. Normality of our data is tested via the Komogorov-Smirnov test for formality.

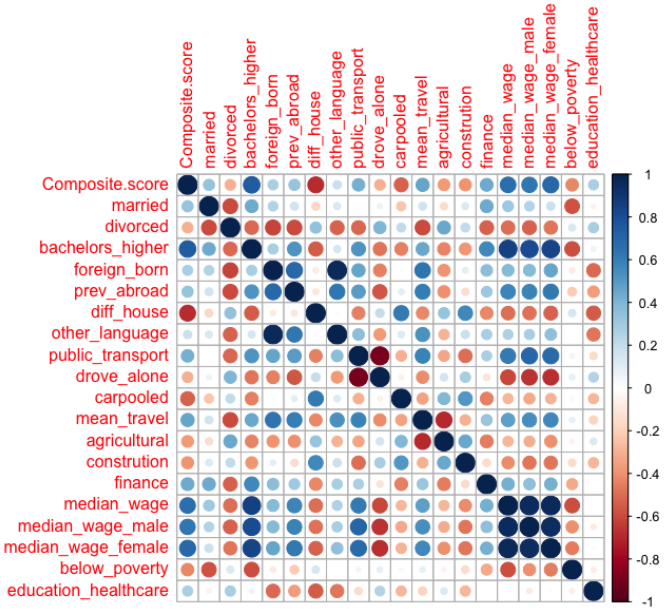
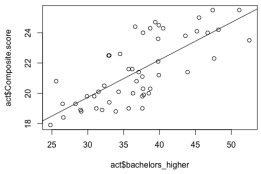
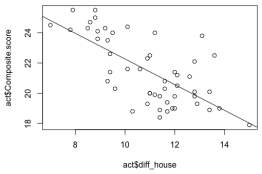


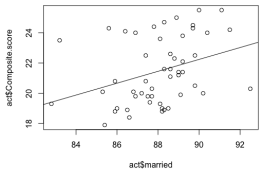
Figure 2: Visual representation of the symmetric correlation matrix between all the different variables in our data.



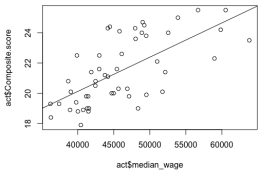
(a) Bachelors or Higher Education



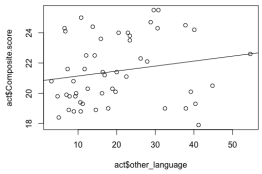
(b) Lived in Different House 1 Year Ago



(c) Married



(d) Median Wage



(e) Other Language Spoken At Home

Figure 3: Relationships between different home socio-economic factors(X-axis) and ACT score(Y-axis).

	Estimated Coefficients (Std. Error)
(Intercept)	12.30^{***}
	(1.21)
bachelors_higher	0.25^{***}
	(0.03)
R^2	0.55
Adj. R^2	0.54
Num. obs.	51

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Education Level

	Estimated Coefficients (Std. Error)
(Intercept)	30.76^{***}
	(1.42)
diff_house	-0.85^{***}
	(0.13)
R^2	0.47
Adj. R^2	0.46
Num. obs.	51

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Different House 1 Year Ago

	Estimated Coefficients (Std. Error)
(Intercept)	-13.12
	(13.89)
married	0.39^{*}
	(0.16)
R^2	0.11
Adj. R^2	0.09
Num. obs.	51

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Marriage Status

	Estimated Coefficients (Std. Error)
(Intercept)	20.81^{***}
	(0.58)
other_language	0.03
	(0.02)
R^2	0.03
Adj. R^2	0.01
Num. obs.	51

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5: Other Language At Home

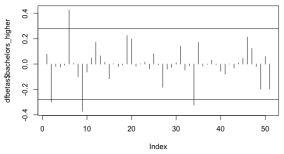
	Estimated Coefficients (Std. Error)
(Intercept)	11.07^{***}
	(1.71)
wage_scaled	2.26^{***}
	(0.37)
R^2	0.44
Adj. R^2	0.42
Num. obs.	51

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

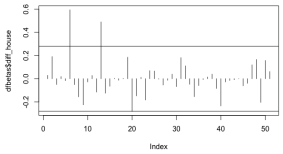
Table 6: Median Wage (Scaled by 10,000)

5.1.2 Influential Observations

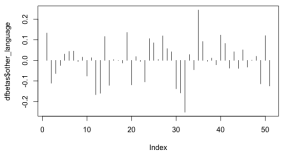
From Figure 4, we see that there are indeed quite a few influential observations in regards to our estimated coefficients. Particularly for higher education and living in a different house 1 year ago. For our bachelors or higher education, we see that there is one large positively influencing data point, and three smaller negatively influencing data points. Thus, one can argue that the outlying positive influencing data point is cancelled out by the three less influential negative outliers, so our bachelors or higher coefficient is not significantly pulled higher or lower. This is not quite the case with our housing covariate, which has two strong positive outliers. This implies that our estimated coefficient is positively influenced by these variables, making it less negative then it might be in reality. However, we know from our regression models that there is a statistically significant negative relationship between living in a different house 1 year ago and ACT score. Thus, these positively influential points are not such a big problem as they only imply that our estimated coefficient may be less negative then it should be. As long as we know that there is a negative relationship our research question is satisfied. For our marriage and wage covariates, the DFBETAS are somewhat less useful here as we know that our full model's estimation of these coefficients is compromised by the negative effects of co-linearity.



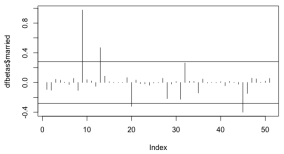
(a) Bachelors or Higher Education



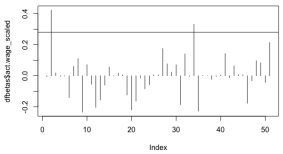
(b) Lived in Different House 1 Year Ago



(c) Other Language



(d) Married



(e) Wages

Figure 4: DFBETAS for each estimated covariate coefficient.

References

[1] Elka Torpey. "Measuring the value of education". In: *U.S. BUREAU OF LABOR STATISTICS* (2018). URL: https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm.

[2] Scott A. Wolla and Jessica Sullivan. "Education, Income, and Wealth". In: *Federal Reserve Bank of St. Louis* (2017). URL: https://research.stlouisfed.org/publications/page1-econ/2017/01/03/education-income-and-wealth/.

[3] Zhonglu Li and Zeqi Qiu. "How does family background affect children's educational achievement? Evidence from Contemporary China". In: *The Journal of Chinese Sociology* (2018). URL: https://journalofchinesesociology.springeropen.com/articles/10.1186/s40711-018-0083-8.

[4] National Center for Education Statistics. *Average ACT scores and percentage of graduates taking the ACT, by state: 2015 and 2019*. data retrieved from National Center for Education Statistics, https://nces.ed.gov/programs/digest/d19/tables/dt19_226.60.asp. 2019.

[5] National Center for Education Statistics. *Education Demographic and Geographic Estimates*. data retrieved from National Center for Education Statistics, https://nces.ed.gov/programs/edge/TableViewer/acsProfile/2019. 2019.