College Scorecard*

Caleb Dowdy, Nicole Navarro, Lance Price, and Eric Voorhis

Is there a significant difference in cost and earnings for the three different school types (Public, Private nonprofit, and Private for-profit) present in the College Scorecard data? In order to measure and quantify these differences a discussion of completeness and model suitability is required. Completeness of the college scorecard data set is assessed in order to provide a hard limit on the utility of said data set and how that impacts the overall reliability of the conclusions drawn. Our statistical analysis consisted of an analysis of variance (ANOVA) and a multiple linear regression model (MLR). Both ANOVA and MLR require that certain assumptions are met in order to guarantee the credibility of the results. The results of our linear modeling indicate that further efforts must be made in order capture diminishing marginal returns of cost in our earnings function and that our earnings function was missing a variable accounting for the talent of the student body.

I. INTRODUCTION

Picking the right school is a very important choice to make. The government has good reason to help students and parents make more informed choices about where students should go to college. It helps the government because there will probably be more productive members of society if they make better choices about their education. These better choices could lead to more positive outcomes of debt repayment, more money cycling through the economic machine, and more taxes being paid to the government. In addition, having more information about colleges can help the government make better choices about which colleges they should invest more money into. All of these motivations are centered around which colleges are most likely to produce higher earning graduates which is important to students and to the government as a whole. This begs the question: how do we define a college's value? One possible definition involves some combination of future earnings and total cost of education.

II. VARIABLES OF INTEREST AND STATEMENT OF THESIS

The variables of interest are median earnings 10 years after enrollment for a given institution and average yearly cost of living plus tuition. We were curious to see whether there were significant differences in these two variables between different school types, so we also looked at the variable named control, which categorizes all schools by whether they are public, private for-profit, or private nonprofit. The goal of this analysis is to detect and quantify a significant difference in cost and earnings across these three types of school if such a difference exists.

* Intended to accompany the Exploratory Data Analysis Presentation.

III. COMPLETENESS

Before we can begin analysis of our data, we must assess the overall completeness of the variables we will be using for the analysis. The original data set is very large, containing data for approximately 7000 schools from 1996 until 2015, but upon inspection it was found that the cost variable we wished to look at was 17.1% complete over the entire data set, and the earnings variable was 18.5% complete. Despite the low levels of completeness, we were still left with approximately 10000 observations due to the large size of our data set originally. In order to decide exactly what analysis was possible with our remaining observations, it was then necessary to take a closer look at the variables to see where the incompleteness stemmed from. Upon closer examination, it was found that cost data was not collected until 2009, but was consistent after that point in time. Median earnings on the other hand exists in the data beginning in 2007, but only exists for the years 2007, 2009, 2011, and 2012. These observations about the data revealed that we possessed enough data to do cross-section or aggregate analysis, but would be unable to do any type of analysis over time.

IV. ANALYSIS

Our analysis of the differences between the types of schools present in the college scorecard dataset consisted of linear modeling; specifically ANOVA and Multiple Linear Regression. These models will be discussed separately in the following subsections.

A. ANOVA

A natural way to analyze the difference among group means is through the use of ANOVA. Several assumptions must be observed in order to correctly perform an ANOVA. Violating any one of these given assumption could result in overestimation and false positives. One of these assumptions requires that variables of interest

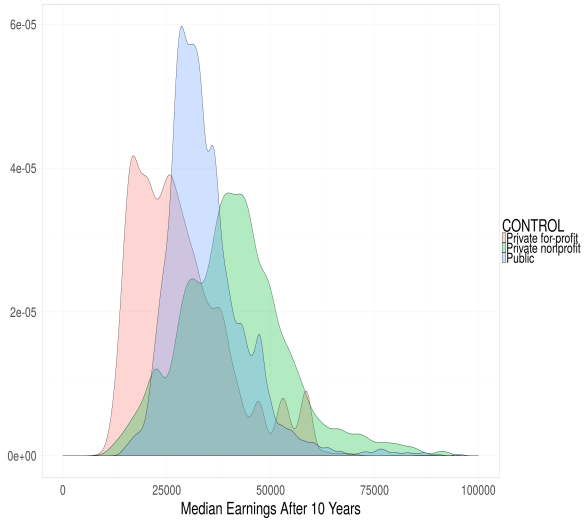
are normally distributed. When we looked at the density plots of the distributions of median earnings for the three school types, the public and private nonprofit school distributions were close to normal distributions (Figure 1). The distribution of median earnings for private for-profit institutions (pink distribution) showed an obvious bimodal distribution. Thus, we only went forward in the ANOVA for the public and private nonprofit schools. The following table shows the results of the ANOVA.

TABLE I. ANOVA of Median Earnings 10 Years Out

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Control	1	1.4E11	1.4E11	835.0	0.00
Residuals	9862	1.6E12	1.6E8		

Table 1 shows that there is a significant difference in the means of median earnings between public and private nonprofit schools, with an F value of 835 (one degree of freedom). We also looked at the distributions of total cost of attendance for the three school types and concluded that ANOVA would not be a suitable way to study the relationship between cost and school type (Figure 2). The distributions for the three school types are either multimodal or have heavy tails. We then moved onto another option for studying the relationship between cost and school type.

FIG. 1. Density Plots of Earnings for Different School Types

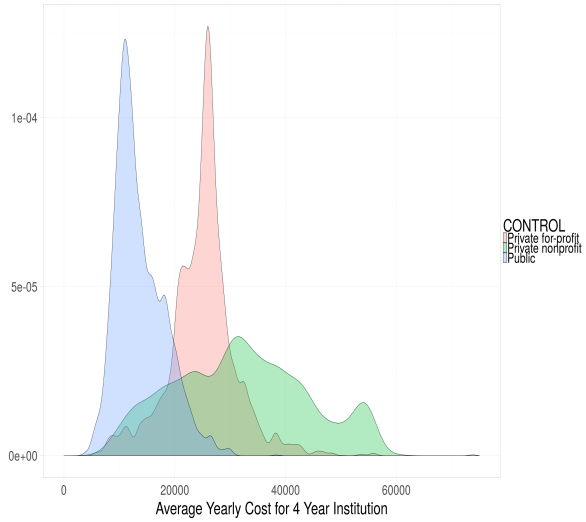


B. Multiple Linear Regression

The goal of this regression is to quantify the relationship that median earnings (10 years after entry) has with respect to cost and type of school. Formally stated:

$$\text{Earnings} = \beta_0 + \beta_1 \text{Cost} + \delta_2 \text{Control} + \epsilon$$

FIG. 2. Density Plots of Cost for Different School Types



Earning := Median earnings for an institution 10 years after entry.

Cost := Average yearly cost of living and tuition for a 4 year institution.

Control := Factor variable for type of institution; private non-profit, private for-profit, or public.

The equation shown above can be thought of as a simple linear regression with the inclusion of a factor variable. Due to the simplicity of this equation it is very limited in it's modeling capabilities.

TABLE II. Multiple Linear Regression[a]

	Dependent variable:
	Earnings
Cost	0.714^{***}
	(0.013)
Control: Private nonprofit	$2,661.872^{***}$
	(289.451)
Control: Public	$9,620.303^{***}$
	(306.431)
Constant	$15,103.640^{***}$
	(398.875)
Observations	7,524
R^2	0.342
Adjusted R^2	0.341
Residual Std. Error	8,919.327 (df = 7520)
F Statistic	$1,300.003^{***}$ (df = 3; 7520)
Note:	*p<0.1; **p<0.05; ***p<0.01

[a] Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu

1. Results

The results in Table 2 indicate that 34% of variation in Earnings are explained by the regression equation. With an F statistic of 1,300.0 we can infer that there is a significant relationship between our independent variables and dependent variable. Coefficients for Cost, Control: Private non-profit, Control: Public, and Constant are all significant at the commonly accepted values of α.

2. Interpreting Cost

Our model suggests that a student at a four year institution can expect $.71 on the dollar spent towards cost of living and tuition, ceterus parabus. Intuitively, the positive relationship between cost and earnings is to be expected because education is an investment, and the greater amount invested suggests a greater return. However, it is also interesting to note that the coefficient for Cost is between zero and one; presumably Earnings is a fraction of Cost because the Cost variable encompasses both the tuition and the cost of living.

3. Interpreting Control

The coefficients of Control: Private nonprofit and Control: Public are measured relative to the missing category (Control: for-profit). Control: Private non-profit has a positive coefficient of 2,661.87 which can be interpreted as the increase in median earnings 10 years after enrollment by attending a private non-profit over a private for-profit. Similarly, the coefficient of 9,620.30 can be interpreted as the increase in median earnings 10 years after enrollment by attending a public institution over a private for-profit institution.

4. Recommendations and Further Research

If a student's main priority is to maximize their earning potential 10 years after enrollment, our model would suggest that they should consider attending a public school and spend as much as possible. For obvious reasons this cannot be the case, there must exist a point at which earnings experiences diminishing marginal returns with respect to cost. We recommend that in further analysis we include a quadratic cost term in order to test our hypothesis regarding diminishing marginal returns. Our model explained roughly 34% of variation in Earnings which leads us to believe that there is lurking variable bias, which could increase our explanatory power. Currently, our model lacks any kind of proxy for the talent present at a given school. A substantial amount of literature suggests a significant positive relationship between an individuals earnings and their IQ. Using a variable for median/mean SAT scores for a given institution would serve as a good proxy for IQ and hopefully increase the overall explanatory power.

V. CONCLUSIONS

After assessing completeness of the data and performing the above analysis, we came to a few conclusions. The first and most jarring conclusion was that the data is very sparse for many of the variables of interest to us. The second main conclusion was that the ANOVA showed a significant difference in median earnings for public and private nonprofit schools. The third main conclusion was that a linear model showed a possible relationship between cost, earnings, and school type.

Future areas to concentrate on would be incorporating average SAT scores into the linear model and calculating the NPV (Net Present Value) of going to each school. This would both help to illuminate what makes a school more likely to produce higher earning graduates to students, the government, and society as a whole.