EDA Project Paper

Chris Leonard, Charlie Edelson, Aaron Niskin, Matthew McAvoy

November 5, 2016

Summary

With rising levels of student debt and decreased earnings, it is important to determine which variable affect median income. Using College Scorecard data from 1996-2015 we investigated the efficacy of levels of degrees conferred, faculty salary, completion rates, family income, and institute tuition per full time equivalent in determining median income ten years post-graduation.

Dataset

The College Scorecard dataset was generated by the United States Department of Education and spans from 1996 to 2015. The dataset was created to assist students and their families in comparing college expenses and outcomes according to personal criteria. The data was reported to College Scorecard by individual institutions and only students receiving financial aid were incorporated into the data.

The dataset consists of 125,703 observations of 1,743 variables. Each observation is at the institution level, where institution is defined by the Integrated PostSecondary Education Data System (IPEDS). Note that this conflicts with the broader institution definition given by the National Center for Education Statistics (NCES) and the Federal Student Aid Office (FSA).

Of the 1,743 variables in the dataset, we were only concerned with 14. Of those 14 we only considered at most 3 at a time and each consideration had over 22,000 observations.

Limitations of the dataset

There are no complete cases in this dataset. This was by design; many categorical attributes were split into multiple separate variables. Additionally the data collection for this dataset began in 1996. This means delayed statistics have a corresponding lag. For instance, median income 10 years after graduation is only available in 2007.

Early updates on this dataset may have been plagued by implementation issues. Institutions may not have been able to generate data on a yearly basis which may explain instances of 2 years condensed onto a single year. All of these lead to inconsistencies in our data.

Research Goal and Motivation

The goal of our investigation was to assess the efficacy of using standard collegiate metrics to predict median income ten years after graduation. We chose median income ten years after graduation not as a metric to assess quality of education, but instead as an economic imperative. Due to rising student debt levels and decreased earnings, it is important to ensure that the return of such an expensive investment is appreciable.

Analysis

Much of the analysis was conducted via exploration, which is out of the scope of this paper. All graphs and plots were produced using `ggplot2` and `qplot` in R. Additionally, all statistics were computed using the R programming language. Given that the dataset has over 1700 variables, we needed an efficient method to

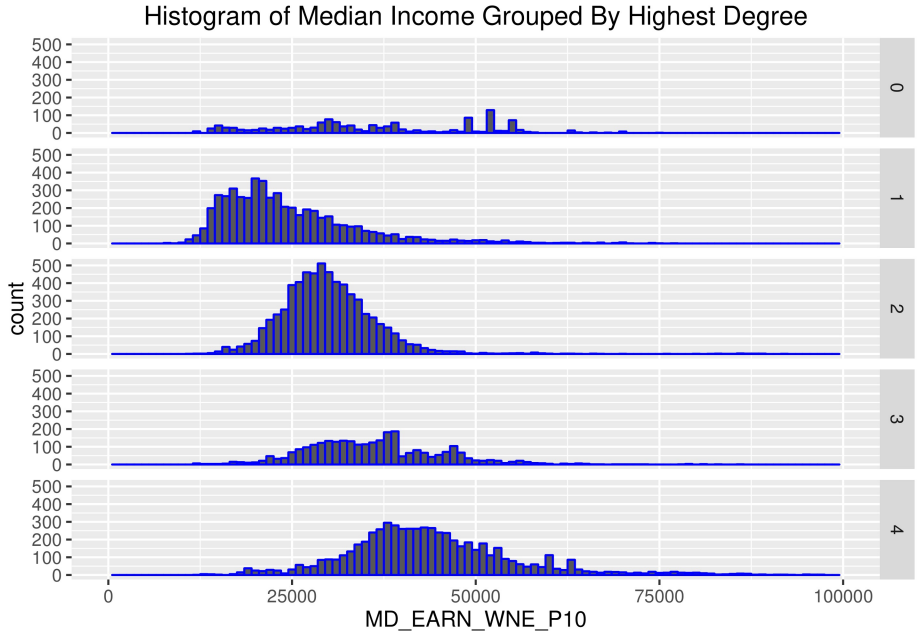


Figure 1: Histrogram showing median income for the grouped by highest degree. Note how distince each distribution is.

eliminate variables. Our method was to focus on several predetermined variables related to school finances, student finances, and institutional degree level.

Results

Institutional Degree Levels

Since bachelor degrees are typically more expensive than certificate programs or associate degrees, it is desirable to know if the investment in a higher level of education will result in a larger return. Furthermore, a school that offers both undergraduate and graduate degrees might have a higher quality of education due to increased resources stemming from the graduate program.

The dataset includes two variables that account for this information: the higheest degree confired (HIGHDEG) by an institute and the predominate degree (PREDDEG). To determine if this had an effect, a one way ANOVE was performed where the reponse variable was median income 10 years out and the explanitory was either highest degree or predominate degree. Both of these ANOVA's were highly significant, with F-statistics over 2000 and p-values $< 2x10^{-16}$. This signifies that grouping by degree significantly explains a large part of the variace. This is evident in Fig 1, where we see the distinct distibution associated with each degree.

Average Faculty Salary and 150% Completion Rate (4 Year)

We chose to only use the 150% completion rate (as opposed to the 100% or 200%) due to the higher correlation with our response variable. For this analysis, we decided to create a categorical variable, which we will denote as `EARNINGS_GT_MD`, which was defined as the logical value of `earnings > median(earnings)`, where `earnings` is the median earnings at 10 years after graduation. This gives us two levels for `EARNINGS_GT_MD`: `TRUE` and `FALSE`. We then plotted the average faculty salary against the 150% completion rate, with the color signifying the value of `EARNINGS_GT_MD` (fig 2). As one can readily see, this is ripe for a logistic regression with an elliptic decision boundary, but unfortunately we could not get the code to work in R by the deadline.

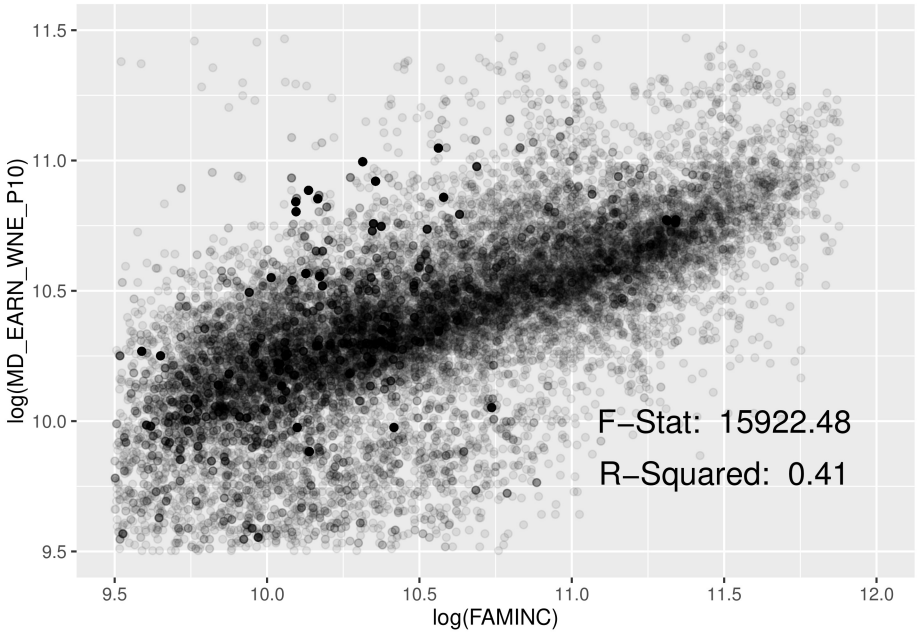
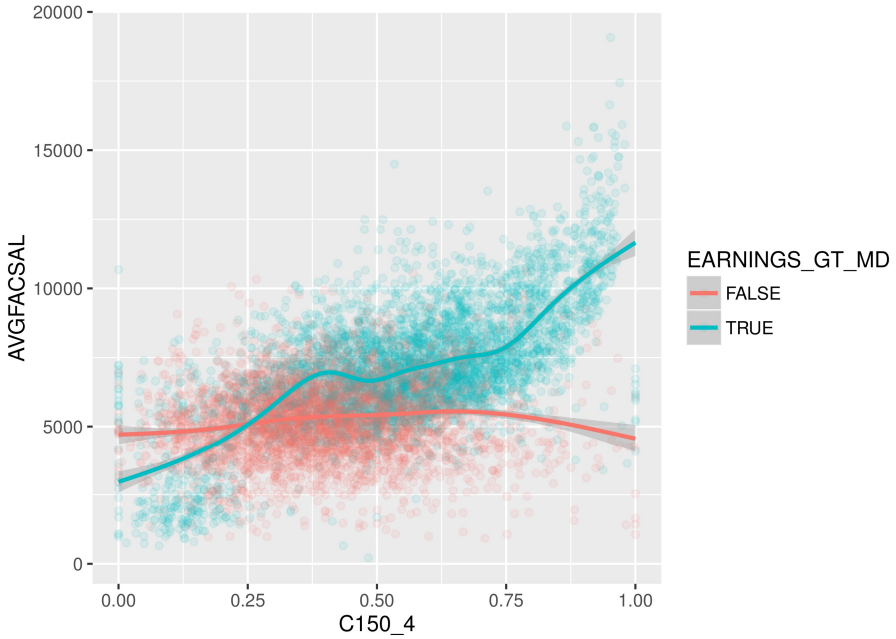


Figure 2: log(Median Earnings) vs log(Median Family Income)

Family Income & Conclusion

The best predictor for median income 10 years after graduation from a four year university, turned out to be family income. Through fitting log(earnings) vs log(fam_inc) with a linear model (fig 3), we were able to determine a likely exponential relationship between earnings and family income. This suggests that the most impactful changes a university can make, in terms of affecting positive change in the median student earnings 10 years after graduation, is to pay the professors more and offer higher degrees.



\begin{figure} \caption{Average Faculty Salary VS 4-Year 150% Completion Rate} \end{figure}