Analyzing the Effect of Financial Aid on Student Academic Success using Multivariate Statistical Models

Michelle Duda

April 29, 2015
List of Figures

3.1 Student aid graphed by year ............................................. 11
3.2 Student aid graphed by sector ........................................... 12
3.3 Yearly student aid dispersion by type ................................. 13
3.4 Total dispersion of student aid by sector ............................. 14
List of Tables

2.1 Sector variable description ........................................ 7

3.1 Percentage of missing data .................................... 15
Acknowledgements

I would like to thank my advisor Dr. Meng Xu for his guidance throughout the project, Dr. Joseph Kolibal for his help with the figures, and the University of New Haven Honors Program for their financial support.
Abstract

Nationwide student loan debt has tripled in the last decade and climbed up by 84% since the latest recession, reaching a historical high of 1.2 trillion U.S. dollars. Student loans have surpassed credit cards and auto loans to become the second biggest source of debt, trailing only behind mortgages. According to The Institute for College Access and Success (TICAS) Project on Student Debt, seven in ten college graduates in 2012 had student loan debt, with an average of $29,400 per borrower. College loan debt not only affects students’ living and academic study during college, but also makes great impact on students’ lives and careers.

As a new round of college debt crisis is prominent, a quantitative understanding of student economics and college data analysis becomes critical. In this thesis, I use student loan data from academic institutions in the United States from the years 2000 to 2010 to study the effect of college loans and aid on students’ academic success. Through a multivariate regression modeling approach, my study aims to illustrate the recent uprise of student aid, to examine various student success measurements, and analyze how students’ aid affects educational attainment and retention. I hope this study would build a quantitative framework for researchers who are interested in college loan issues and provide college guidance to students as well as financial information to school educators and administrators.
Chapter 1

Introduction

Student aid is aimed to provide students with financial support for their higher education studies. However, with the recent financial crisis and increase of college tuition, student aid has become more of a burden than a relief for many students, especially for those who face graduation and have just begun their professional careers. According to the Student Loan Debt Clock, at the time of this writing, student loan debt is closing 1.26 trillion dollars and increasing at a rapid rate of $2,854 per second [13]. According to The Institute for College Access and Success (TICAS) Project on Student Debt, seven in ten college graduates in 2012 had student loan debt, with an average of $29,400 per borrower.

The impact of student aid on students success in college has been studied in literature. Dynarski (1999), using a mixed-effect linear model, predicted that the increase of federal grant amount would improve the attendance and completion rate. Specifically, she studied the Social Security Student Benefit Program which gave financial aid to 18-22 year olds whose parents were deceased or were receiving social security benefits because of disability or re-
tirement. Congress voted to eliminate this program in 1981 which caused a huge change for those benefiting from this program before its demise. Dynarski concluded that a $1,000 increase in the grant aid for an eligible person increases the time to achieve educational goals by about 0.16 years and the probability of attending college by about 4% [9].

Dowd and Coury (2006) used survey data from the National Center for Education Statistics (NCES) to study the effect of financial aid, primarily federal loans, on persistence and associate degree attainment for community college students admitted in 1990. Data for a follow-up group in 1994 was also analyzed. In their logistic models, students academic performances and their racial, family, and financial status were included as predictor variables to predict the enrollment rate after two years in college and the percentage of associates degree attained after five years. They found that, statistically, loans had a negative effect on persistence and no effect on degree attainment [7].

Bozick (2007) studied persistence of first year college students based on various economic constraints. He proposed the following two questions: Are students from families with limited economic resources more likely than their affluent peers to hold jobs and to live at home during their first year of college? Does working and living at home during the first year of college impede academic persistence? To answer these questions, Bozick used bivariate cross-tabulations, a series of logistic regression models with weights. Bozick found that students from high income families had the lowest rate of employment. However, students from lowest income families did not have the highest rate of employment [2]. Bozick suggested this may be because these students receive more need-based aid or because students in this category are usually
ethnic minorities who may have trouble finding jobs because they are viewed negatively in the work environment. Bozick also found that students in the lowest income bracket were more likely to live at home with their parents and students in the highest income status were more likely to live on campus in their first year of school.

Besides works mentioned above, other various authors have studied the effects of loans and other financial variables on college student lives and education using specific cohorts (McPherson and Schapiro 1991 [17]; Volkwein et al. 1998 [14]; Kim 2004 [15]; Manthei and Gilmore 2005 [16]; Dowd and Coury 2006 [7]; Whalen and Shelley 2010 [19]; Johnson et al. 2013 [12]). However, most of the previous works relied on logistic regression techniques and focused on a specific response variable (e.g. enrollment rate, completion rate, etc.). The effect of financial aid on college student academic successes using multiple measures has not been fully studied.

The current work uses the 2000-2010 longitudinal data sets from the Integrated Postsecondary Education Data System (IPEDS) of the National Center for Education Statistics (NCES) to study the impact of financial aid on US undergraduate students education outcomes. Multiple logistic regressions and multivariate analysis are utilized to analyze the statistical significance of these effects. I will compare the results obtained from logistic regression using a single success measure with those from multivariate regressions using multiple success measures. Our study differs from the previous literature in that multivariate analysis allows one to examine the effects of multiple response variables (here as academic success measures) simultaneously.
Chapter 2

Methodology

2.1 Data and Variables

Data used in this work are retrieved from the National Center for Education Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS). The original table [10] (available from Delta cost project database) lists detailed school finance, enrollment, and staffing data from 1978 to 2010. For the purpose of this study, we are interested in variables that measure students academic success and financial aid. We used the more recent data from 2000 to 2010 in our analysis.

In order to determine a relationship between student finance and student academic success, we first need to select variables to quantify these broad terms. For the financial variables, a search through the data dictionary with keywords including “loans”, “tuition”, and “aid” was done. For academic success variables, “graduation rate”, “degrees earned” and “retention rate” were searched as keywords.

Specifically, the following were selected as student\textsuperscript{1} financial aid variables:

\textsuperscript{1}In this study, a student is defined to be a full-time, first-time degree-seeking undergraduate student.
• *fed grant num*: the number of students receiving federal grants, including grants and educational assistance funds,

• *fed grant pct*: the ratio of *fed grant num* over the total number of students attending the institution,

• *fed grant avg amount*: the average amount of federal grants received by students including Title IV Pell Grants, Supplemental Educational Opportunity Grants, and other grants or educational assistance funds provided through the federal government or any federal agencies,

• *state grant num*: the number of students receiving state or local grants in the form of grants, scholarships and waivers provided through state and local student aid programs,

• *state grant avg amount*: the average amount of state or local grants in the form of grants, scholarships, and waivers received by students provided through state and local student aid programs,

• *state grant pct*: the ratio of *state grant num* over the total number of students attending the institution,

• *inst grant num*: the number of students receiving institutional grants in the form of scholarships and fellowships provided through the institution or individual departments within the institution,

• *inst grant pct*: the ratio of *inst grant num* over the total number of students attending the institution,

• *loan num*: the number of students receiving student loans which must be repaid to the lending institution the loan was borrowed from, including all Title IV subsidized and unsubsidized loans and all institutionally-
and privately-sponsored loans, not including PLUS or other loans made directly to parents,

- **loan_pct**: the ratio of `loan_num` over the total number of students attending the institution,

- **loan_avg_amount**: the average amount of loans which must be repaid to the lending institution the loan was borrowed from, including all Title IV subsidized and unsubsidized loans and all institutionally- and privately-sponsored loans, not including PLUS or other loans made directly to parents made to students,

- **net_student_tuition**: total amount of students’ net tuition not including Pell, Federal, State, and Local grants.

For a given institution, in a given year, the total amount of financial aid of each type was calculated as the product of the number and the average amount in the corresponding type. The following were selected as student academic success variables:

- **bachelordegrees**: the total number of bachelor degrees conferred by students within at least four but no more than five years of college-level work,

- **grad_rate_150_n**: the number of students graduating within 150 percent of normal time, which is defined as the time necessary to complete all requirements for a degree or certificate as set by the institution,

- **grad_rate_150_p**: the ratio of `grad_rate_150_n` over the total number of students attending the institution,

- **ftretention_rate**: the ratio of the number of students re-enrolled at the institution during the current fall term as part-time or full-time since
admission in the previous fall term over the number of students admitted in the previous fall term.

To examine the trends and effects of financial aids for different schools, we analyzed the national data for four major types of schools through use of the variable *sector_revised*, a categorical variable defined by the following:

<table>
<thead>
<tr>
<th>Sector Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>public, 4 years or more</td>
</tr>
<tr>
<td>2</td>
<td>private not for profit, 4 years or more</td>
</tr>
<tr>
<td>3</td>
<td>private for profit, 4 years or more</td>
</tr>
<tr>
<td>4</td>
<td>public, 2 to 4 years</td>
</tr>
</tbody>
</table>

Table 2.1: Sector variable description

### 2.2 Models

Logistic regression is widely applied in educational data researches (Peng et al. 2002 [4]). In a logistic model, the response variable represents the binary outcome of a Bernoulli experiment: failure or success. The expected response function $E(y)$, which gives the probability of success $\pi$ in each trial (here as one student) is defined as,

$$E(y) = \pi = \frac{1}{1 + e^{-X'\beta}}$$  \hfill (2.1)

The maximum likelihood method is used to estimate the logit function $\eta = \ln \left( \frac{\pi}{1 - \pi} \right)$ and then to calculate the model parameters $\beta$s. In the current work, I will build a multiple logistic regression (MLR) model using average student aid amounts from different sources, including average student loan, average tuition fee, academic year and school type, as predictor variables $X$. Full-time student retention rate, full-time student graduation rate within 150% of normal time, and percentage of bachelor degree awarded are used as the
response variables separately. This analysis tests the effects of financial aids on student successes at the level of each individual student. The Wald test is used to infer the statistical significance of each predictor variable in a given model. Likelihood ratio test is used to examine the overall significance of the regression, by comparing the existing model with a reduced model of constant probability of success (Montgomery et al. 2012 [6]), for which,

$$E(y) = \pi = \frac{1}{1 + e^{-\beta_0}} \quad (2.2)$$

In a multivariate multiple regression (MMR) model, “multivariate” means multiple response variables and “multiple” means multiple predictor variables. MMRs allow one to construct statistical hypotheses of regression coefficients across equations for different response variables simultaneously. This is the most compelling reason to use MMRs other than ordinary linear regressions (OLR), which deal with one response variable at a time (UCLA: Statistical Consulting Group [11] 2015).

Unlike the MLR, when building the MMR, we used the total amount of tuition, loans, and grants from different sources as predictor variables, and total number of students who received bachelor degree, total number of students who graduated within 150% of normal completion time, and the total number of students who returned after the first year at school as response variables. Unlike the MLR testing aid effects at the student level, MMR examines the effects of financial aid on academic successes at the school level.
Chapter 3

Results

3.1 Descriptive Statistics

The total federal grant amount summed over all colleges of the nation (Fig. 3.1a), increased steadily from 2000 to 2004, and was constant between 2004 to 2006, then climbed rapidly from 2007 to 2010 two fold. The total amount of loans (Fig. 3.1d) increased at a steady rate until 2007, when it began to increase rapidly and doubled in 2010. Total institutional grant aid (Fig. 3.1c) exhibited accelerating growth during 2000 to 2010, whereas state grant aid (Fig. 3.1b) and tuition (Fig. 3.1e) increased linearly. Overall, I saw a dramatic increase in federally supplied grant aid and loan amounts in the most recent years while the other variables increased with the same rate from 2000 to 2010.

Total federal grants of each type of school, pictured in Figure 3.2a, followed the same pattern as that of the overall amount in Fig 3.1a. State grant (Fig. 3.2b), however, varied by sector. For sectors 4 and 1, state grant increased over time, but state grants awarded to institutions in sector 1 increased more rapidly. Sectors 2 and 3 increased very little or not at all from 2000 to 2010. Total institutional grant aid (Fig. 3.3c) was very small for sector
and sector 4 and graphically, these values look insignificant in respect to the remaining two sectors. Sectors 1 and 2 both increased quadratically, however, sector 2 increased most rapidly. Total loan aid (Fig. 3.3d) followed a similar pattern as total federal grant aid; it increased slowly, plateaued, and increased rapidly for a few years before beginning what looks like a plateau. Total tuition (Fig. 3.3e) followed a linear pattern, with sectors 1, 2, and 3 increasing. Sector 4, however, initially followed the linear increase but experienced a quick decrease only for 2002, and then began decreasing again after 2008. Overall sector 1 received the most financial aid, specifically these schools received the most federal grant, state grant, and loan aid. Sector 2, received the most institutional grant aid and had the highest tuition. From our analysis, we see the two types of schools that receive the most aid are the schools with the highest tuition.

As seen in Figure 3.3, the amount of each type of aid including, federal, institutional, loan, and state, stays relatively the same with a slight change in 2010 (Fig. 3.3f). Loan aid encompasses approximately 40% of total aid. Institutional grant aid amounts to approximately 29% of all aid. Approximately 20% of total aid is federal grant aid. Lastly, approximately 11% of all aid is state grant aid. We see a decrease in state grant and an increase in federal grant.

Depicted in Figure 3.4, The sector with the greatest amount of total aid is private not for profit, 4 years or more followed by public, 4 years or more, followed by public, 2 to 4 years, and lastly, private for profit, 4 years or more. Over the 10 years, sectors 3 and 4 fluctuate and end up receiving equal amounts of aid.
Figure 3.1: Student aid graphed by year
Figure 3.2: Student aid graphed by sector
Figure 3.3: Yearly student aid dispersion by type
Figure 3.4: Total dispersion of student aid by sector
3.2 Logistic Regression

The variables outlined in Table 3.1 were used in our logistic regression analysis.

Three full logistic regression models, one for each response variables $y_1, y_2, y_3$

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Percentage of Missing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>fed_grant_avg_amount($x_1$)</td>
<td>18%</td>
</tr>
<tr>
<td>state_grant_avg_amount($x_2$)</td>
<td>25%</td>
</tr>
<tr>
<td>inst_grant_avg_amount($x_3$)</td>
<td>24%</td>
</tr>
<tr>
<td>loan_avg_amount($x_4$)</td>
<td>26%</td>
</tr>
<tr>
<td>academic_year($x_5$)</td>
<td>0%</td>
</tr>
<tr>
<td>sector_revised($x_6$)</td>
<td>0%</td>
</tr>
<tr>
<td>tuition_avg($x_7$)</td>
<td>19%</td>
</tr>
<tr>
<td>grad_rate_150_p($y_1$)</td>
<td>38%</td>
</tr>
<tr>
<td>bachelor_degrees_pct($y_2$)</td>
<td>28%</td>
</tr>
<tr>
<td>ftretention_rate($y_3$)</td>
<td>50%</td>
</tr>
</tbody>
</table>

Table 3.1: Percentage of missing data

...were first created using $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ as the predictor variables. For each response variable, a single “best” MLR model was selected using backward elimination and the likelihood ratio test. Here, “best” means that a model described the data as well as the corresponding full model. Results for the three best MLRs were described as follows.

When graduation rate was used as the response variable, $inst_grant_avg_amount$, academic_year, tuition_avg showed statistically positive effects on the graduation rate. For every additional dollar increase in $inst_grant_avg_amount$, the odds of graduation increased by $8.1 \times 10^{-3}\%$. For every additional dollar increase in $tuition_avg$, the odds of graduation increased by $1.7 \times 10^{-3}\%$. In the 10-year span from 2000 to 2010, the graduation rate increased significantly. In addition, sectors 2 and 4 differed significantly from sector 1 with respect to the graduation rate.
When the percentage of bachelor degree was used as the response variable, the effects of \textit{inst\_grant\_avg\_amount} and \textit{tuition\_avg} were statistically significant and positive. For every additional dollar increase in \textit{inst\_grant\_avg\_amount}, the odds of obtaining a bachelor degree increased by $1.5 \times 10^{-3}\%$. For every additional dollar increase in \textit{tuition\_avg}, the odds of obtaining a bachelor degree increased by $1.0 \times 10^{-3}\%$. Sectors 3 and 4 differed significantly from sector 1 with respect to the percentage of bachelor degree.

When retention rate was used as the response variable, the effects of \textit{inst\_grant\_avg\_amount} and \textit{tuition\_avg} were statistically significant and positive. The effect of \textit{academicyear} was statistically significant and negative. For every additional dollar increase in \textit{inst\_grant\_avg\_amount}, the odds of a student staying in school after the first year increased by $6.5 \times 10^{-3}\%$. For every additional dollar increase in \textit{tuition\_avg}, the odds of a student staying in school after the first year increased by $1.8 \times 10^{-3}\%$. The retention rate dropped from 2000 to 2010. Sectors 2, 3, and 4 all differed from sector 1 with respect to the retention rate.

3.3 Multivariate Analysis

We used multivariate multiple regression (MMR) to study the relationship between student financial aid and academic success. MMR differed from MLR in that the former used overall count measures for each school while the latter treated the variables as percentages or average per student. Therefore MMR gave a school-level analysis of the relationship. MMR showed that the amount of grants, loan, and tuition all had significantly positive effects on each response variable, with one exception. Total federal grant amount had
significantly negative effect on the total number of college graduates.

Multivariate analysis of variance (MANOVA) tested the effect of each financial variable on the three response variables simultaneously (see Methods). MANOVA function in R showed that the amount of grants, loan, and tuition all exhibited significant effects on graduation.
Chapter 4

Conclusions and Discussions

I conclude that institutional grant and tuition fees significantly increased the student academic success measured by graduation rate, retention rate, and percentage of bachelor degree earned per student. The positive effect of tuition fees on student success, while counter-intuitive, may be accounted for by the large amount of institutional grant in corresponding institutions. It would be interesting to develop a new financial aid measure that incorporate the correlation between institutional grant and tuition, and re-examine the statistical effects of predictor variables studied in this thesis.

Surprisingly, multiple logistic regressions showed that college loans were not significant contributors to the student academic success or failure. Multivariate regression showed that college loans had positive effects on all student success measures. However, this effect may be statistical artifact due to the multicollinearity among predictor variables (i.e. college loans and financial aid were increasing simultaneously over the years).

Further studies should examine the relationship between financial aid and
loan and student success at the state level. It is reasonable to speculate that
different state policies can explain the results from the corresponding model
analysis. It is also interesting to include student work study variables into our
model, so that the impact of self-finance by students on their academic success
can be examined.

Recently, various student loan pay-back proposals were examined (Dy-
narski and Kreisman 2013 [8]). It will be interesting to study how the new
proposals would affect student’s academic performance and their education
attainments as examined here.

Despite the substantial proportion of missing data, we omitted the missing
values of each variable when building the MLRs and MMR and did not make
any corrections to the data. We did not use any data imputation techniques
to estimate the missing values as it may cause unexpected biases and errors
in the data and in our results (Allison 2000 [1]). Multivariate probit models
(Cappellari and Jenkins 2003 [3]), which have the features of both MLR and
MMR, are difficult to interpret and therefore, were not selected as one of our
models.
Bibliography


