Which colleges are worth it?

Predicting debt-to-earnings ratios using regression analysis

Omair A. Khan / *oak3@psu.edu*

March 18, 2016

Contents

Abstract

Student loan debt in the United States is at an all time high. However, data from the Bureau of Labor Statistics show that college graduates earn much more than those with high-school diplomas. Using a variety of regression methods (ridge, lasso, principal component, and partial least squares), this paper attempts to model a school's return on investment by purely economic means: the debt-to-earnings ratio. Schools with a low debt-to-earnings ratio imply that either students take on less debt attending or earn proportionally higher salaries than the debt they take on. Lasso regression provides the best fit with the lowest test error and highest coefficient of determination. We find that certain universities have the best ROI and art schools and for-profit colleges tend to be worse choices. Penn State's main campus outperforms the branch campuses in terms of the ratio. This analysis is intended to serve as a tool for future college students to help guide their decision making on which school to attend.

Introduction

With student loan debt in the United States exceeding $1.2 trillion and 17% of borrowers in default (The Hechinger Report 2015), the high cost of a college education has been heatedly debated by presidential candidates this year. Almost 71% of bachelor's degree recepients graduated with a student loan in 2015, compared to less than 50% two decades ago (Sparshott 2015). The average class of 2015 graduate will have to pay back over $35,000 according to an analysis at the Urban Institute (Baum and Johnson 2015). Graduates who received Pell Grants were much more likely to borrow and to borrow more (College Access & Success 2014). With such unprecedentedly high student loan debt afflicting our nation's young adults, one wonders whether college is worth the cost. The U.S. Bureau of Labor Statistics has shown that full-time workers with at least a bachelor's degree earn almost twice as much as those with only a high-school diploma (Sparshott 2015). If a young adult is interested in having a higher potential income, how can they choose a college with the best return on investment?

Using information from the College Scorecard dataset (U.S. Department of Education 2015), this paper attempts to answer this important question using historic data and statistical modeling. The debt-to-earnings ratio will be predicted for each school based on a number of variables in this regression analysis. The purpose of this modeling is to provide statistically literate adults a purely economical tool to evaluate a school's return on investment to advise future students and try to get at the question, "which colleges are worth it?" We will focus on the following three Research & Statistical Questions.

Research questions

1. How can we best predict the return on investment for all colleges and universities in the United States?
2. What types of schools have a poor return on investment?
3. Using The Pennsylvania State University as a case study for schools with branch campuses, is it ever advantageous to study at a sattelite (Commonwealth) campus to save money?

Statistical questions

1. Using a variety of regression techniques, which model has the lowest test error and highest coefficient of determination?
2. What types of schools have a very high debt-to-earnings ratio?
3. What is the relationship between the debt-to-earnings ratio for the University Park campus compared to the sattelite campuses?

We will assume that the relationship between debt-to-earnings ratio and multiple predictors such as ethnic makeup, SAT scores, location, and admission rate (among others) can be modeled in a linear fashion. Four

types of regression will be conducted, ridge, lasso, principal component, and partial least squares, to find one model that has the best fit to the test data and the lowest prediction error rate.

Methods

Preparation

Select data (see **Table 1**) from years 2009 and 2011 of predominantly bachelor's degree granting schools were imported from a SQLite database into R. The variables chosen for inclusion were selected using domain knowledge of the research questions rather than heuristically. The median debt-to-earnings ratio was calculated by dividing the median debt of the graduates by the median earnings six years after entry. A number of data types were changed to standardize naming and organize string, numeric, and factor data. Data from 2009 was used to create the training set (with missing observations removed) and data from 2011 was used to create the test set (with missing observations removed).

Table 1: Variables used to train and predict debt-to-earnings ratios

Variable (SQL)	Variable (R)	Description
ADM_RATE	AdmRate	Admission rate
AVGFACSAL	FacultySalary	Average faculty salary
C150_4	CompletionRate	Completion rate for first-time, full-time students
CDR2	Default2	Two-year cohort default rate
CONTROL	CollegeType	College type (public, private non-profit, or private for-profit)
COSTT4_A	Cost	Average cost of attendance
GRAD_DEBT_MDN	MedianDebtGrad	Median debt of completers
INSTM	College	Institution name
md_earn_wne_p6	Earnings6	Median earnings of students working and not enrolled 6 years after entry
md_earn_wne_p8	Earnings8	Median earnings of students working and not enrolled 8 years after entry
md_earn_wne_p10	Earnings10	Median earnings of students working and not enrolled 10 years after entry
PCTFLOAN	FSFLoans	Percent of all federal undergraduate students receiving a federal student loan
PPTUG_EF	PartTime	Share of undergraduate students who are part-time
RPY_1YR_RT	Repayment1	Fraction of repayment cohort that has not defaulted, and with loan balances that have declined one year since entering repayment
RPY_3YR_RT	Repayment3	Fraction of repayment cohort that has not defaulted, and with loan balances that have declined three years since entering repayment
RPY_1YR_N	NSRepayment1	Number of students in the 1-year repayment rate cohort
SATMTMID	Math	Midpoint of SAT scores at the institution (Math)
SATVRMID	Verbal	Midpoint of SAT scores at the institution (Verbal)
SATWRMID	Writing	Midpoint of SAT scores at the institution (Writing)
STABBR	abb	State
UGDS	UndergradEnrollment	Enrollment of undergraduate degree-seeking students
UGDS_WHITE	WEnroll	Share of undergraduate students who are white
UGDS_BLACK	BEnroll	Share of undergraduate students who are black
UGDS_HISP	HEnroll	Share of undergraduate students who are Hispanic
UGDS_ASIAN	AEnroll	Share of undergraduate students who are Asian

Variable (SQL)	Variable (R)	Description
WDRAW_DEBT_MDN	MedianDebtNGrad	Median debt for students who have not completed
YEAR	Year	Year of data (only 2009 and 2011 were kept)
GRAD_DEBT_MDN / md_earn_wne_p6	DebtToEarn	Median debt-to-earnings ratio

Regression

The seed value for the Mersene Twister pseudorandom number generator is first set to $x_0 = 1234$ to ensure consistency and reproducibility. The first type of regression performed is ridge regression. The value for the tuning parameter λ is found via cross-validation. The training model is then computed after which the testing set is used to predict debt-to-earnings ratios. R^2 and test error are finally calculated. Lasso regression proceeds the same way.

Next, principal component regression is used to fit the training data. A validation plot is used to find the number of components with the minimum mean square error of prediction (MSEP). The debt-to-earnings ratio is then predicted for the test data using the determined number of components and finally R^2 and test error are calculated. The final model, using partial least squares regression, is found using the same procedure as principal component regression.

Statistical programming

R version 3.2.4 (R Core Team 2015) was used to perform all statistical analyses on the College Scorecard dataset (U.S. Department of Education 2015). Observations from 2009 (with missing values removed) were used to predict the debt-to-earnings ratios from 2011 as these two years had the greatest number of non-missing data.

The data was coded in an SQL database which was interfaced with R using the RSQLite package (Wickham, James, and Falcon 2014). Other R packages used include dplyr for data manipulation and tidying (Wickham and Francois 2015); Amelia for working with missing data (Honaker, King, and Blackwell 2011); glmnet for ridge and lasso regression (Friedman, Hastie, and Tibshirani 2010); pls for partial least squares and principal component regression (Mevik, Wehrens, and Liland 2015); and corrgram, choroplethr, and Hmisc (Wright 2015; Lamstein 2014; Harrell Jr 2015) for data visualization.

Exploratory data analysis

After creating the training dataset from 2009, missing observations were removed to reduce the amount of imputation necessary. The missing data is visualized in **Figure 1**, where each row of pixels is a separate observation and yellow implies missing data at the specified variable. **Figure 2** shows a correlogram of the various predictors. Variable pairs with deeper reds indicate stronger positive correlations while pairs with deeper blue indicate stronger negative correlations. Rather than removing variables with strong correlations to prevent over-fitting, we decided to leave them in as three of the four regression methods are also forms of dimension reduction.

Statistical analysis

This paper's analysis focuses on four methods of linear model selection: ridge, lasso, principal component, and partial least squares regression. For the sake of simplicity, we assume that the relationship between the debt-to-earnings ratio and the other variables can be modeled using linear methods. There is evidence of

non-linearity in some predictors as shown by a post-hoc ANOVA for non-parametric effects and this should be followed up in the future with spline and generalized additive models.

Ridge regression

Ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

where $\lambda \geq 0$ is a tuning parameter determined by cross-validation.

Ridge regression with cross-validation (see **Figure 3**) yielded a model with R^2 of 0.301 and a test error of 0.0370.

Lasso regression

Lasso regression is similar to ridge regression except that it uses an ℓ_1 penalty instead of an ℓ_2 penalty, thus effectively performing variable selection. The lasso coefficients $\hat{\beta}_\lambda^L$ minimize the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Lasso has the benefit of yielding sparse models which usually outperform those of ridge regression.

Lasso regression with cross-validation (see **Figure 4**) yielded a model with R^2 of 0.318 and a test error of 0.0361. This is the best performing model of the four. The coefficients in **Table 2** can be interpreted as the amount of each variable needed to increase the debt-to-earnings ratio by one point. Note that some of the coefficients are missing; this is due to either singularity of data or the dimension reduction feature of using an ℓ_1 penalty.

Table 2: Regression coefficients from best (lasso) model

Predictor	Coefficient	Predictor	Coefficient
(Intercept)	0.3825	abbNJ	-0.0154
FSFLoans	0.2907	abbNM	.
CollegeTypeNonprofit	-0.0224	abbNV	.
CollegeTypePublic	.	abbNY	-0.0106
CompletionRate	-0.0016	abbOH	0.0503
Default2	2.3394	abbOK	0.0438
Cost	0	abbOR	0.0751
abbAR	0.1422	abbPA	0.0239
abbAZ	.	abbRI	0.0196
abbCA	.	abbSC	0.1634
abbCO	-0.0022	abbSD	0.0140
abbCT	-0.0106	abbTN	0.0275
abbDC	0.0321	abbTX	.
abbDE	-0.2214	abbUT	-0.0277
abbFL	0.0011	abbVA	-0.0050
abbGA	.	abbVI	-0.1307
abbHI	0.0921	abbVT	0.0126
abbIA	.	abbWA	0.0007

Predictor	Coefficient	Predictor	Coefficient
abbID	0.1468	abbWI	0.0422
abbIL	-0.0827	abbWV	0.0318
abbIN	0.0370	FacultySalary	0
abbKS	0.0021	Repayment1	0.0034
abbKY	0.0298	Repayment3	-0.0030
abbLA	0.0692	NSRepayment1	.
abbMA	-0.0080	PLanguage	0.1542
abbMD	-0.0489	AdmRate	0.0261
abbME	0.0391	PartTime	-0.0493
abbMI	0.0885	Math	-0.0016
abbMN	0.0517	WEnroll	-0.0035
abbMO	0.0880	BEnroll	0.1249
abbMS	.	HEnroll	.
abbMT	.	AEnroll	-0.1288
abbNC	0.0853	Verbal	0.0010
abbND	0.0502	Writing	0.0004
abbNE	-0.0514	UndergradEnrollment	0
abbNH	-0.0525		

Principal component regression

Principal component regression is another dimension reduction method. The original predictors $X_1, X_2, \ldots, X_p$ are first transformed into $M < p$ linear combinations represented by $Z_1, Z_2, \ldots, Z_M$ where $Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$ for some constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ $(m = 1, \ldots, M)$. The selection of the constants ϕ_{jm} is obtained using an unsupervised approach. The regression model is then fit with these transformed predictors.

Principal component regression yielded a model with R^2 of 0.172 and a test error of 0.0439. This is the worst performing model of the four. This is a surprising find and is possibly explained by non-linearity in the relationship of debt-to-earnings ratio and other variables.

Partial least squares regression

Partial least squares regression is also a dimension reduction method. It follows the same procedure as principal component regression except that the selection of the constants ϕ_{jm} (and therefore the selection of Z_m) is obtained using a supervised approach.

Partial least squares regression yielded a model with R^2 of 0.300 and a test error of 0.0371.

Discussion

Table 3 shows a summary of R^2 and test error values for each model. Answering Research & Statistical Question 1, the lasso model outperforms the other regression methods with the lowest test error of 0.0361. A graphical summary of the true and predicted debt-to-earnings ratios by state for public schools (chosen to make comparisons between states more uniform) is shown in **Figures 5-7**. The model was unable to predict values for the following states due to singularity of observations after removing missing data: Alaska, District of Columbia, Iowa, Kansas, Mississippi, Missouri, Nebraska, Tennessee, and Wyoming. We notice that the R^2 values for each model are still very poor. This is likely explained by the presence of non-linearity (against our assumption) in certain predictors.

Table 3: Summary of prediction results for each model

Method	R^2	Test error
Ridge	0.3011656	0.03703867
Lasso	0.3182226	0.03613464
PCR	0.1715142	0.04391028
PLS	0.3002878	0.03708519

Ideally, the results in **Table 3** should be compared to baseline values obtained from ordinary least squares regression. We encountered many errors when trying to create a simple linear model, particularly the number of predicted observations did not match the number of test observations. Because of this, we could not calculate R^2 or test error values for the OLS model and therefore excluded it all together. This was discussed with Dr. Maggie Niu (The Pennsylvania State University, Department of Statistics) but debugging did not yield better results.

To address Research & Statistical Question 2, we first look at the best predicted schools with the lowest debt-to-earnings ratios (**Table 4**). It is important to remember when looking at this list that many observations were excluded due to missing data (**Figure 1**) so the table actually shows the top 10 schools with the highest debt-to-earnings ratios that had complete data.

Table 4: Top 10 schools with the best predicted ROI

Rank (Best)	School
1	University of Illinois at Urbana-Champaign
2	Harvard University
3	Claremont McKenna College
4	Massachusetts Institute of Technology
5	Princeton University
6	University of Delaware
7	Northwestern University
8	California Institute of Technology
9	Yale University
10	Stanford University

The top 10 schools with the highest debt-to-earnings ratios (after removing missing values) are shown in **Table 5**. Schools 1-7 have a debt-to-earnings ratio that is greater than 1. This means that students from these colleges and universities have more debt than earnings six years after entry.

Table 5: Top 10 schools with the worst predicted ROI

Rank (Worst)	School
1	University of Arkansas at Pine Bluff
2	Livingstone College
3	Edward Waters College
4	Philander Smith College
5	Morehouse College
6	Lincoln University of Pennsylvania
7	Tennesee Temple University
8	Bethune-Cookman University
9	Fisk University
10	Kansas City Art Institute

Perusing the list of true debt-to-earnings ratios in 2011 (due to removal of missing observations), we see that certain types of post-secondary schools stand out as having the worst return on investment. As a group, the various Art Institute campuses (art and design schools); DeVry College campuses (for-profit technical schools); and University of Phoenix campuses (for-profit schools with little-to-no recognition) have the highest debt-to-earnings ratios, making these colleges a poor investment for a student's future.

Finally, to answer Research & Statistical Question 3, we find that The Pennsylvania State University University Park campus has the lowest debt-to-earnings ratio (0.533) compared to the Commonwealth campuses ($0.534 \geq \text{DTE} \geq 0.686$). This may be due to a variety of factors such as better networking, better education, and larger sample size of students at the University Park campus. The statistics do not take into consideration students in the 2+2 plan who start at a Commonwealth campus and transfer to University Park for their junior and senior years.

Conclusion

This analysis has shown that there are indeed certain factors which make a college worth the investment. SAT scores and faculty salary had the lowest p-values of the signficant predictors. Despite the limitations due to non-linearity, we have provided a strong step towards building a useful model to evaluate debt-to-earnings ratios for post-secondary education. We hope that this analysis can help future students make informed decisions on which schools to attend.

Figures

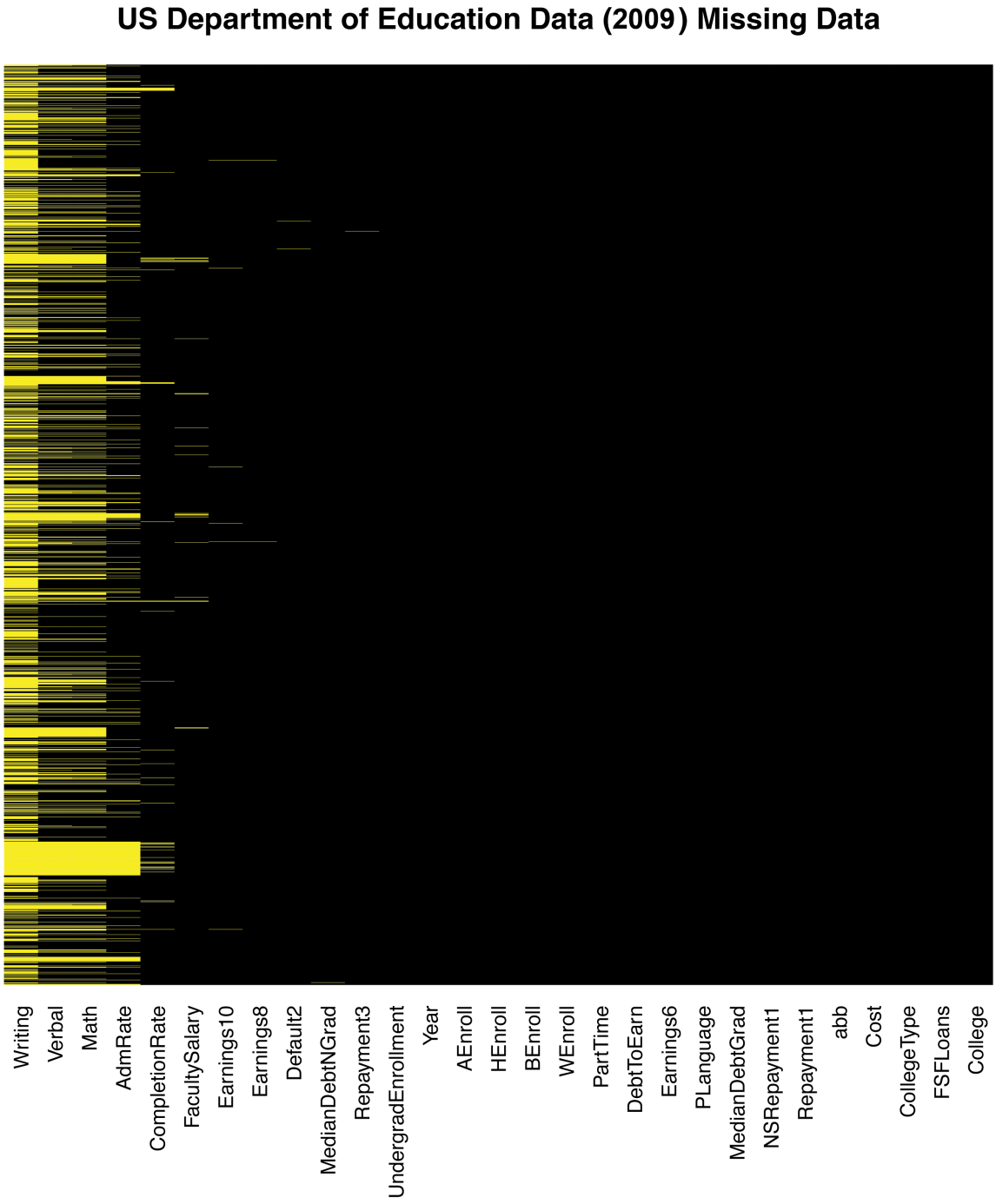

Figure 1: Missing data in original training dataset

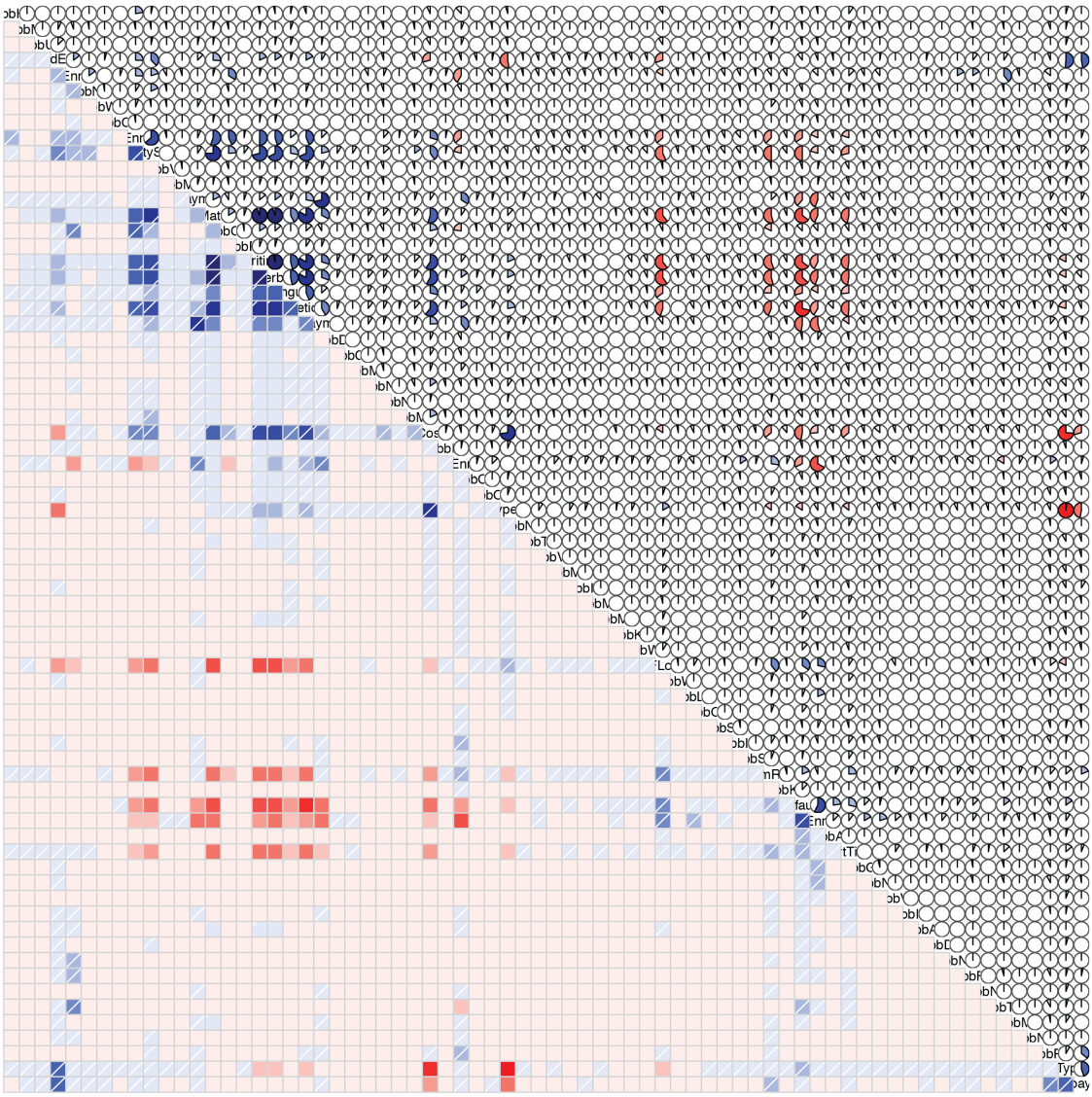

Figure 2: Correlogram of included predictors

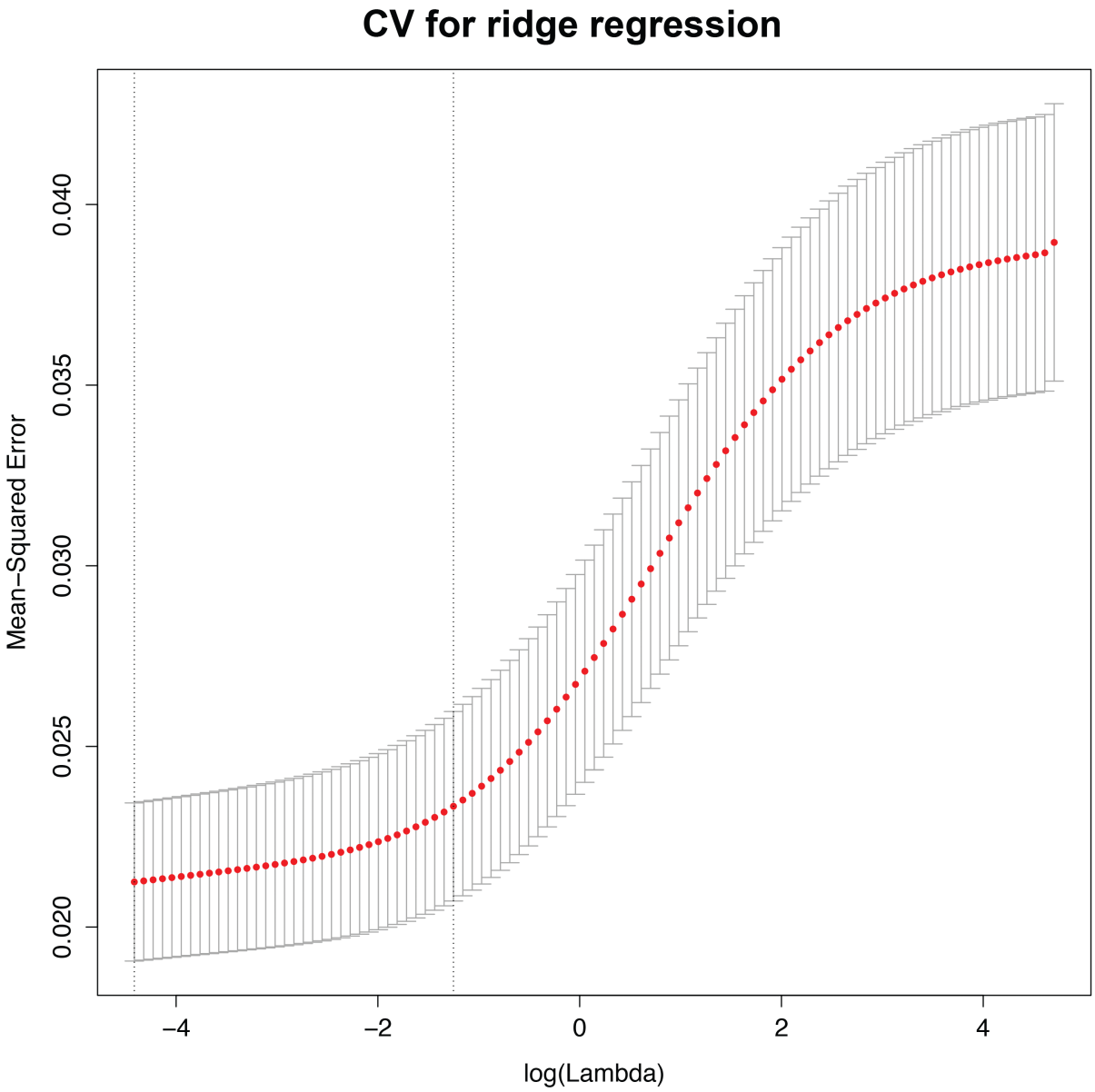

Figure 3: Cross-validation for lambda in ridge regression

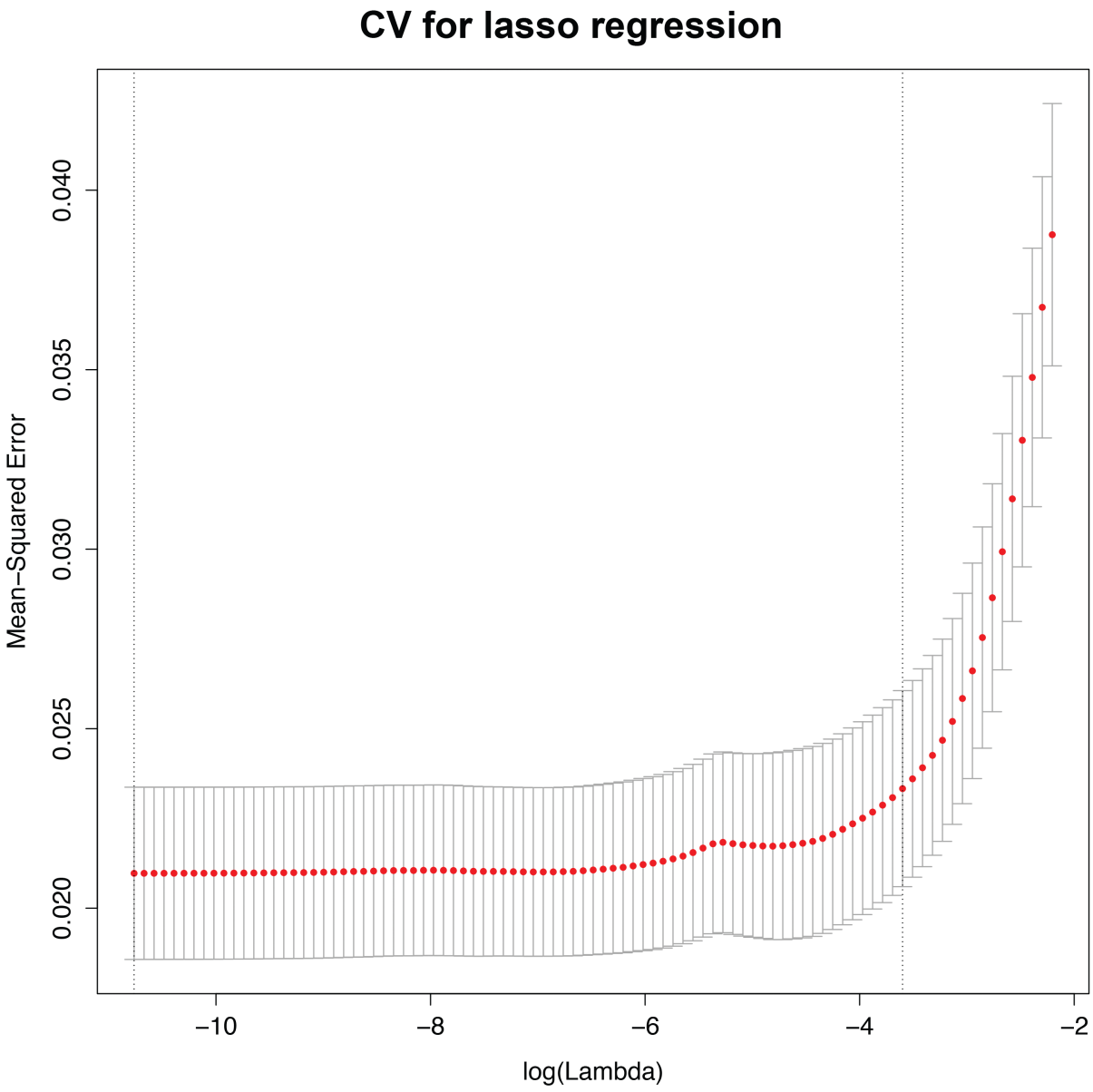

Figure 4: Cross-validation for lambda in lasso regression

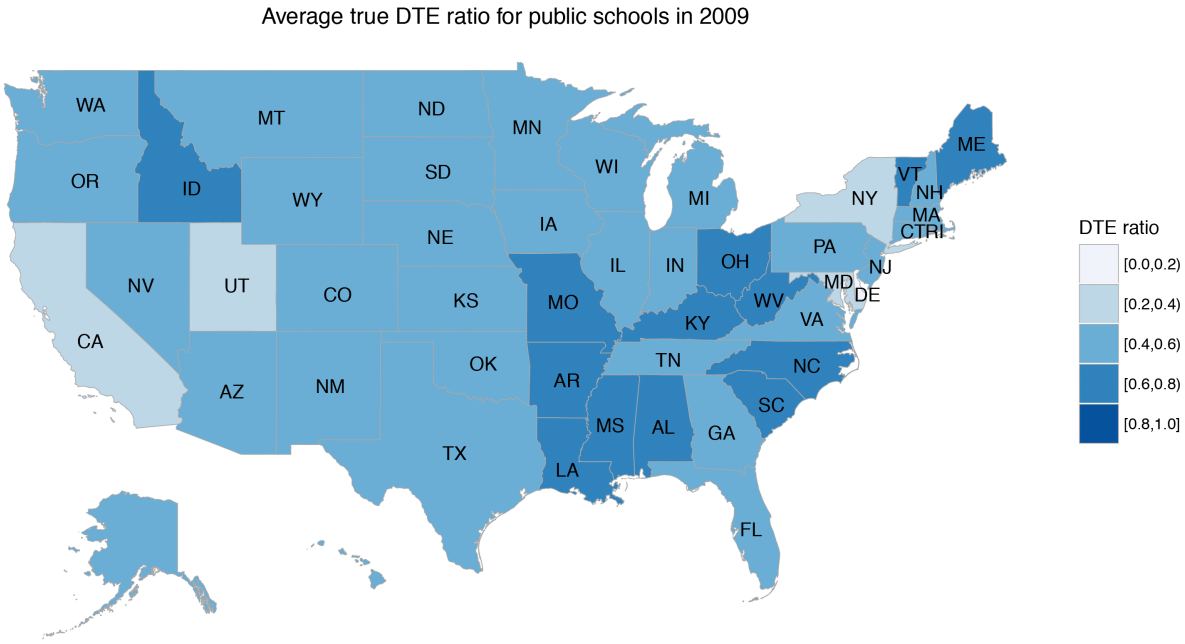

Figure 5: Average true debt-to-earnings ratio for public schools in 2009

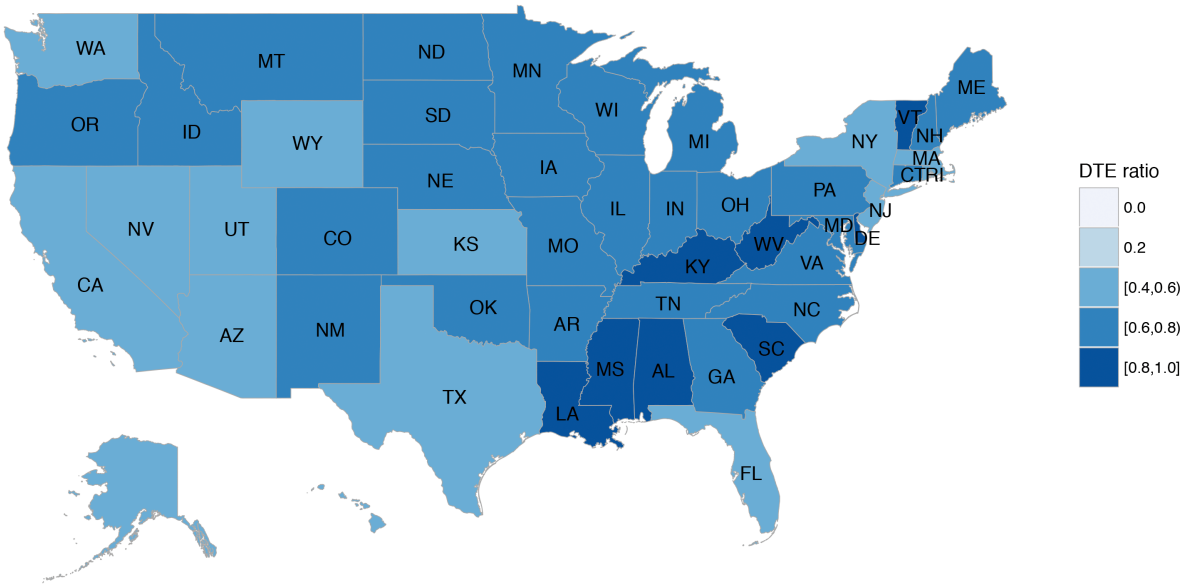

Figure 6: Average true debt-to-earnings ratio for public schools in 2011

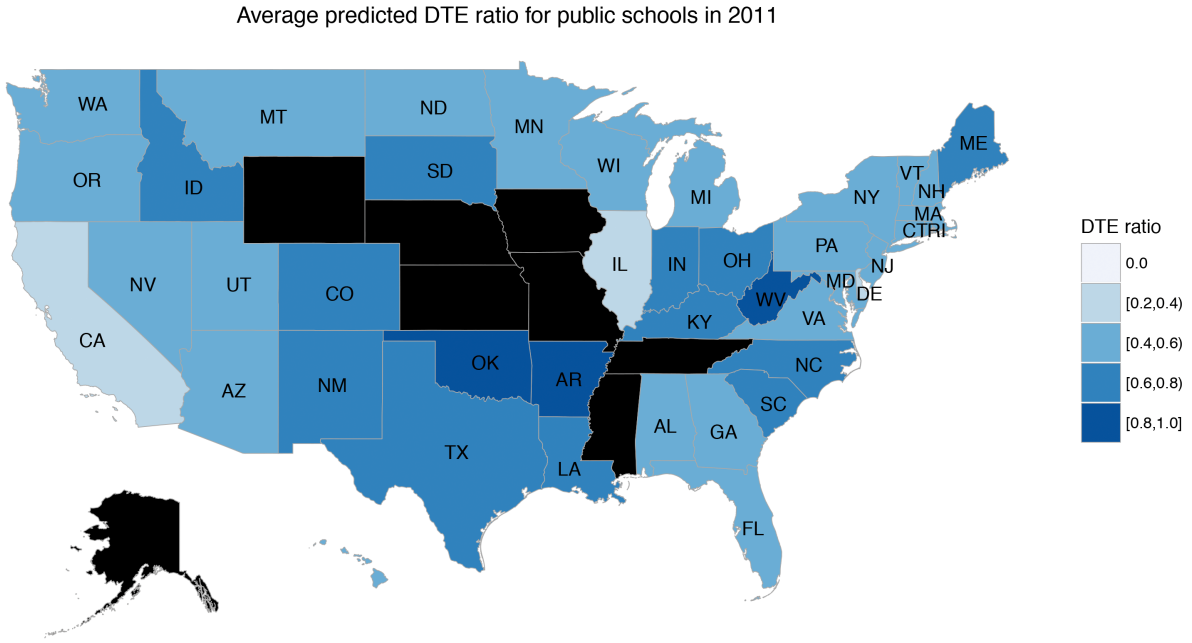

Figure 7: Average predicted debt-to-earnings ratio for public schools in 2011

References

Baum, Sandy, and Martha C. Johnson. 2015. "Student Debt: Who Borrows Most? What Lies Ahead?" *Urban Institute*. http://www.urban.org/research/publication/student-debt-who-borrows-most-what-lies-ahead/view/full_report.

College Access & Success, The Institute for. 2014. "Quick Facts About Student Debt." http://ticas.org/sites/default/files/pub_files/Debt_Facts_and_Sources.pdf.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. http://www.jstatsoft.org/v33/i01/.

Harrell Jr, Frank E. 2015. *Hmisc: Harrell Miscellaneous*. https://CRAN.R-project.org/package=Hmisc.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7): 1–47. http://www.jstatsoft.org/v45/i07/.

Lamstein, Ari. 2014. *ChoroplethrMaps: Contains Maps Used by the Choroplethr Package*. https://CRAN.R-project.org/package=choroplethrMaps.

Mevik, Bjørn-Helge, Ron Wehrens, and Kristian Hovde Liland. 2015. *Pls: Partial Least Squares and Principal Component Regression*. https://CRAN.R-project.org/package=pls.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sparshott, Jeffrey. 2015. "Congratulations, Class of 2015. You're the Most Indebted Ever (for Now)." *Wall Street Journal*. http://blogs.wsj.com/economics/2015/05/08/congratulations-class-of-2015-youre-the-most-indebted-ever-for-now/.

The Hechinger Report. 2015. "Heaviest Debt Burdens Fall on 3 Types of Students." *U.S. News and World Report*. http://www.usnews.com/news/articles/2015/06/08/heaviest-college-debt-burdens-fall-on-3-types-of-students.

U.S. Department of Education. 2015. "College Scorecard Data (Version 1.7.7)." https://collegescorecard.ed.gov/data/.

Wickham, Hadley, and Romain Francois. 2015. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, David A. James, and Seth Falcon. 2014. *RSQLite: SQLite Interface for R*. https://CRAN.R-project.org/package=RSQLite.

Wright, Kevin. 2015. *Corrgram: Plot a Correlogram*. https://CRAN.R-project.org/package=corrgram.

Appendix: R code

```
## Title: Predicting Student Loan Debt to Earnings Ratios: A Regression Analysis
## Author: Omair Khan
## Date: 2015-03-09

## Working Directory
setwd("~/Dropbox/Documents/School/2016 Spring/STAT 581")

## Load libraries
library(RSQLite) # To connect to database
library(choroplethr) # For US map
library(choroplethrMaps)
library(Hmisc) # For Choropleth cuts
library(Amelia) # For showing NA's in graph
library(dplyr) # For working with data frames
library(glmnet) # For lasso regression
library(pls) # For PCR and PLS regression
library(corrgram) # For visualizing correlations

## Connect to database
db <- dbConnect(dbDriver("SQLite"), "~/Downloads/output/database.sqlite")
dbGetQuery(db, "PRAGMA temp_store=2;")

## Load state data
data("state.regions")

#_____________________________________________________________________________
# Retrieve data from scorecard

## 2009 data
Loans2009 <- dbGetQuery(db, "
                        SELECT  INSTNM College,
                        PCTFLOAN FSFLoans,
                        CONTROL CollegeType,
                        C150_4 CompletionRate,
                        CDR2 Default2,
                        COSTT4_A Cost,
                        STABBR abb,
                        AVGFACSAL FacultySalary,
                        RPY_1YR_RT Repayment1,
                        RPY_3YR_RT Repayment3,
                        RPY_1YR_N NSRepayment1,
                        GRAD_DEBT_MDN MedianDebtGrad,
                        PCIP16 PLanguage,
                        ADM_RATE AdmRate,
                        WDRAW_DEBT_MDN MedianDebtNGrad,
                        md_earn_wne_p6 Earnings6,
                        md_earn_wne_p8 Earnings8,
                        md_earn_wne_p10 Earnings10,
                        GRAD_DEBT_MDN/md_earn_wne_p6 DebtToEarn,
```

```
                        PPTUG_EF PartTime,
                        SATMTMID Math,
                        UGDS_WHITE WEnroll,
                        UGDS_BLACK BEnroll,
                        UGDS_HISP HEnroll,
                        UGDS_ASIAN AEnroll,
                        Year,
                        SATVRMID Verbal,
                        SATWRMID Writing,
                        UGDS UndergradEnrollment
                        FROM Scorecard
                        WHERE Year=2009
                        AND PREDDEG='Predominantly bachelor''s-degree granting'
                        AND md_earn_wne_p6 NOT LIKE 'NA'
                        AND GRAD_DEBT_MDN NOT LIKE 'NA'
                        AND GRAD_DEBT_MDN/md_earn_wne_p6 NOT LIKE 'Inf'
                        AND Cost NOT LIKE 'NA'
                        ORDER BY INSTNM ASC")

## 2011 data
Loans2011 <- dbGetQuery(db, "
                        SELECT  INSTNM College,
                        PCTFLOAN FSFLoans,
                        CONTROL CollegeType,
                        C150_4 CompletionRate,
                        CDR2 Default2,
                        COSTT4_A Cost,
                        STABBR abb,
                        AVGFACSAL FacultySalary,
                        RPY_1YR_RT Repayment1,
                        RPY_3YR_RT Repayment3,
                        RPY_1YR_N NSRepayment1,
                        GRAD_DEBT_MDN MedianDebtGrad,
                        PCIP16 PLanguage,
                        ADM_RATE AdmRate,
                        WDRAW_DEBT_MDN MedianDebtNGrad,
                        md_earn_wne_p6 Earnings6,
                        md_earn_wne_p8 Earnings8,
                        md_earn_wne_p10 Earnings10,
                        GRAD_DEBT_MDN/md_earn_wne_p6 DebtToEarn,
                        PPTUG_EF PartTime,
                        SATMTMID Math,
                        UGDS_WHITE WEnroll,
                        UGDS_BLACK BEnroll,
                        UGDS_HISP HEnroll,
                        UGDS_ASIAN AEnroll,
                        Year,
                        SATVRMID Verbal,
                        SATWRMID Writing,
                        UGDS UndergradEnrollment
                        FROM Scorecard
                        WHERE Year=2011
                        AND PREDDEG='Predominantly bachelor''s-degree granting'
```

```
                        AND md_earn_wne_p6 NOT LIKE 'NA'
                        AND GRAD_DEBT_MDN NOT LIKE 'NA'
                        AND GRAD_DEBT_MDN/md_earn_wne_p6 NOT LIKE 'Inf'
                        AND Cost NOT LIKE 'NA'
                        ORDER BY INSTNM ASC")

#_______________________________________________________________________________
# Regression preparation

## Rename CollegeType to prevent errors because of spacing
Loans2009$CollegeType[Loans2009$CollegeType == "Private nonprofit"] <- "Nonprofit"
Loans2009$CollegeType[Loans2009$CollegeType == "Private for-profit"] <- "For-profit"
Loans2011$CollegeType[Loans2011$CollegeType == "Private nonprofit"] <- "Nonprofit"
Loans2011$CollegeType[Loans2011$CollegeType == "Private for-profit"] <- "For-profit"

## Create test data
df.test <- select(Loans2011, College, DebtToEarn, FSFLoans, CollegeType, CompletionRate,
                  Default2, Cost, abb, FacultySalary, Repayment1, Repayment3,
                  NSRepayment1, PLanguage, AdmRate, PartTime, Math,
                  WEnroll, BEnroll, HEnroll, AEnroll, Verbal, Writing,
                  UndergradEnrollment)

df.test <- na.omit(df.test)

x.test <- model.matrix(DebtToEarn ~ . -College, df.test)[,-1]
y.test <- df.test$DebtToEarn

## Set up test.avg for R2 calculation
test.avg <- mean(y.test)

## Calculate DTE ratio for each state in 2009
df1 <- group_by(Loans2009, CollegeType, abb)
DTE <- summarize(df1, value = mean(DebtToEarn))

## Change variables to appropriate class
Loans2009$Cost <- as.numeric(Loans2009$Cost)
Loans2009$abb <- as.factor(Loans2009$abb)
Loans2009$FacultySalary <- as.numeric(Loans2009$FacultySalary)
Loans2009$NSRepayment1 <- as.numeric(Loans2009$NSRepayment1)
Loans2009$UndergradEnrollment <- as.numeric(Loans2009$UndergradEnrollment)

## Select variables for training data set
df.train <- select(Loans2009, College, DebtToEarn, FSFLoans, CollegeType, CompletionRate,
                   Default2, Cost, abb, FacultySalary, Repayment1, Repayment3,
                   NSRepayment1, PLanguage, AdmRate, PartTime, Math,
                   WEnroll, BEnroll, HEnroll, AEnroll, Verbal, Writing,
                   UndergradEnrollment)

## Remove observations with missing data
df.train <- na.omit(df.train)

## Convert data frame to matrix
x.train <- model.matrix(DebtToEarn ~ . -College, df.train)[,-1]
```

```r
y.train <- df.train$DebtToEarn

## Remove states that are not common between 2009 and 2011: AL, GU, PR, WY
x.train <- x.train[, -c(7, 17, 46, 59)]

## Create data frame for MSE and R2
results <- matrix(data = NA, nrow = 4, ncol = 2)
results <- as.data.frame(results)
colnames(results) <- c("MSE", "R2")
rownames(results) <- c("Ridge", "Lasso", "PCR", "PLS")

#_____________________________________________________________________________
# Exploratory data analysis

## Visualizing missing data
missmap(Loans2009, main = "US Department of Education Data (2011) Missing Data",
        col = c("yellow", "black"), legend = FALSE)

## Visualizing correlation
corrgram(x.test, order = TRUE, lower.panel = panel.shade, upper.panel = panel.pie,
         text.panel = panel.txt)

# ++++++++++++++++++++++++++
# flattenCorrMatrix
# ++++++++++++++++++++++++++
# cormat : matrix of the correlation coefficients
# pmat : matrix of the correlation p-values
flattenCorrMatrix <- function(cormat, pmat){
  ut <- upper.tri(cormat)
  data.frame(
    row <- rownames(cormat)[row(cormat)[ut]],
    column <- rownames(cormat)[col(cormat)[ut]],
    cor <- (cormat)[ut],
    p <- pmat[ut]
  )
}

res <- rcorr(x.test)
correlations <- flattenCorrMatrix(res$r, res$P)

#_____________________________________________________________________________
# Ridge regression

grid <- 10^seq(10, -2, length = 100)

set.seed(1234)

## Cross validation to find optimal lambda
cv.out <- cv.glmnet(x.train, y.train, alpha = 0)
plot(cv.out, main = "CV for ridge regression")
bestLambdaRidge <- cv.out$lambda.min

## Model with all observations
```

```
ridge.fit <- glmnet(x.train, y.train, alpha = 0, lambda = grid, thresh = 1e-12)

## Predict for 2011 data
ridge.pred = predict(ridge.fit, s = bestLambdaRidge,
                     newx = x.test)

## Calculate and assign MSE and R2 for all school types
results[1, 1] <- mean((y.test - ridge.pred)^2)
results[1, 2] = 1 - mean((y.test - ridge.pred)^2) / mean((y.test - test.avg)^2)

## Find coefficients
ridge.coef <- predict(ridge.fit, type = "coefficients",
                      s = bestLambdaRidge, newx = x.test)

#_____________________________________________________________________________
# Lasso regression

set.seed(1234)

## Cross validation to find optimal lambda
cv.out <- cv.glmnet(x.train, y.train, alpha = 1, family = "gaussian")
plot(cv.out, main = "CV for lasso regression")
bestLambdaLasso = cv.out$lambda.min

## Model with all observations
lasso.fit <- glmnet(x.train, y.train, alpha = 1, family = "gaussian")

## Predict for 2011 data
lasso.pred <- predict(lasso.fit, newx = x.test, type = "response",
                      s = bestLambdaLasso, exact = TRUE)

## Calculate and assign MSE and R2 for all school types
results[2, 1] <- mean((y.test - lasso.pred)^2)
results[2, 2] <- 1 - mean((y.test - lasso.pred)^2) / mean((y.test - test.avg)^2)

## Find coefficients
lasso.coef <- predict(lasso.fit, newx = x.test, type = "coefficients",
                      s = bestLambdaLasso, exact = TRUE)

#_____________________________________________________________________________
# PCR regression

set.seed(1234)
pcr.fit <- pcr(y.train ~ x.train, validation = "CV")
summary(pcr.fit)

validationplot(pcr.fit, val.type = "MSEP")
MSEP(pcr.fit)

pcr.pred <- predict(pcr.fit, newdata = x.test, ncomp = 47)
pcr.pred <- pcr.pred[1:683]

results[3, 1] <- mean((y.test - pcr.pred)^2)
```

```
results[3, 2] <- 1 - mean((y.test - pcr.pred)^2) / mean((y.test - test.avg)^2)

#_______________________________________________________________________________
# PLS regression

set.seed(1234)
pls.fit <- plsr(y.train ~ x.train, validation = "CV")
summary(pls.fit)

validationplot(pls.fit, val.type = "MSEP")
MSEP(pls.fit)

pls.pred <- predict(pls.fit, newdata = x.test, ncomp = 27)
pls.pred <- pls.pred[1:683]

results[4, 1]   <- mean((y.test - pls.pred)^2)
results[4, 2] <- 1 - mean((y.test - pls.pred)^2) / mean((y.test - test.avg)^2)

#_______________________________________________________________________________
# Calculations for maps

## True debt-to-earnings ratio for 2009
DTEpublic2009 <- filter(DTE, CollegeType == "Public")
DTEpublic2009 <- inner_join(x = DTEpublic2009, y = state.regions, by = "abb")
DTEpublic2009 <- select(DTEpublic2009, CollegeType, value, region)

## True debt-to-earnings ratio for 2011
df2 <- group_by(Loans2011, CollegeType, abb)
DTE2011 <- summarize(df2, value = mean(DebtToEarn))

DTEpublic2011 = filter(DTE2011, CollegeType == "Public")
DTEpublic2011 = inner_join(x = DTEpublic2011, y = state.regions, by = "abb")
DTEpublic2011 = select(DTEpublic2011, CollegeType, value, region)

## Lasso (best model) predicted debt-to-earnings ratio for 2011
pred2011 <- cbind(df.test$College, df.test$CollegeType, df.test$abb, lasso.pred)
colnames(pred2011) <- c("College", "CollegeType", "abb", "DebtToEarn")
pred2011 <- as.data.frame(pred2011, stringsAsFactors = FALSE)
pred2011$DebtToEarn <- as.numeric(pred2011$DebtToEarn)

df3 <- group_by(pred2011, CollegeType, abb)
DTEpred <- summarize(df3, value = mean(DebtToEarn))

DTEpred <- left_join(x = DTEpred, y = state.regions, by = "abb")
DTEpred <- select(DTEpred, CollegeType, value, region)

DTEpredpublic <- filter(DTEpred, CollegeType == "Public")

#_______________________________________________________________________________
# Map visualizations

## True debt-to-earnings ratio for 2009
DTEpublic2009$value=cut2(DTEpublic2009$value, cuts = c(0, 0.2, 0.4, 0.6, 0.8, 1))
```

```
print(state_choropleth(DTEpublic2009,
                       title="Average true DTE ratio for public schools in 2009",
                       num_colors = 5,
                       legend="DTE ratio"))

## True debt-to-earnings ratio for 2011
DTEpublic2011$value=cut2(DTEpublic2011$value,cuts=c(0, 0.2, 0.4, 0.6, 0.8, 1))

print(state_choropleth(DTEpublic2011,
                       title="Average true DTE ratio for public schools in 2011",
                       num_colors = 5,
                       legend="DTE ratio"))

## Lasso (best model) predicted debt-to-earnings ratio for 2011
DTEpredpublic$value=cut2(DTEpredpublic$value,cuts=c(0, 0.2, 0.4, 0.6, 0.8, 1))

print(state_choropleth(DTEpredpublic,
                       title="Average predicted DTE ratio for public schools in 2011",
                       num_colors = 5,
                       legend="DTE ratio"))
```