College Admissions Findings

Erica Wong, Elly Wang, Lily Li, Bryana Gutierrez

December 5, 2016

1 Objective

Our goal for this project is to help our client, who are group of administrators trying to make their school more competitive. In order to achieve this goal, we will be making a shiny app that will help us pin point which areas our client can improve on to lower their admissions rates. In this app, we will be able to give our clients a list of schools that have similar attributes to them. Additionally, we will be able to compare our client's school's characteristics with other schools that are similar, but have a lower admission rate to help see what areas they can improve on to lower the admission rate. Finally, we will be able to predict our client's new admission rates based on a regression equation that we have derived.

2 Introduction

In order to be able to predict the our client's admission rates and find out which variables are impactful, we first had to come up with a model via regression. We decided to look at many different methods of regression so that we would be able to compare mean-squared errors, and eventually pick the model that produces the lowest mean squared error. We looked at ordinary least squares (OLS), ridge, lasso, principal components, and partial least squares regression. With ordinary least squares we did model selection and cross validation to find the best OLS model. Ridge and lasso regression are shrinkage methods and principal components and partial least squares are reduction methods. The purpose of comparing the different types of methods is so we can find the best model possible. We want to be able to compare the models produced by all the methods so we will have prediction accuracy and proper model interpretation.

We want to be able to compare different models and do model selection because it will help us to narrow down our model to the variables that are impactful and correlated to the admission rate. Having less variables is helped because it there will be less models to interpret and the model will be easier to understand because we know that each variable is significant in some way. These methods all bring some value to our model, but in different ways. So, we want to compare all of them with each other and see how the model produced using each method differs from one another in order to make the most well rounded decision.

3 Data

For our project, we are working with data from *College Scorecard*. A link to this website can be found here. *College Scorecard* was created to help students find schools that best fit their needs and goals. The way that the raw information is presented is split apart by academic year and contains many different variables. So, the first thing that we had to do was select the columns of variables that we wanted to look at further for the years we wanted and make a new file for that. We called this file **colleges.csv**.

In **colleges.csv**, we have many different columns. The columns and what they stand for are:

- OPEID: Office of Post Secondary Education ID of Institution
- INSTNM: Institution Name
- CITY: City
- STABBR: State Abbreviation
- ZIP: Zip Code
- UGDS: Undergraduate Student Size
- GDS_WHITE: Percentage of White Undergraduates
- UGDS_BLACK: Percentage of Black Undergraduates
- UGDS_HISP: Percentage of Hispanic Undergraduates
- UGDS_ASIAN: Percentage of Asian Undergraduates
- UGDS_AIAN: Percentage of American Indian/Alaska Native Undergraduates
- UGDS_NHPI: Percentage of Native Hawaiian/Pacific Islander Undergraduates
- UGDS_2MOR: Percentage of Two or More Races Undergraduates
- UGDS_NRA: Percentage of Non-Resident Alien Undergraduates
- UGDS_UNKN: Percentage of Unknown Race Undergraduates
- AGE_ENTRY: Average Age of Undergraduates when Entering Institution
- UGDS_WOMEN: Percentage of Women Undergraduates
- MARRIED: Percentage of Married Undergraduates
- FIRST_GEN: Percentage of First Generation College Students
- FAMINC: Average Family Income

- MN_EARN_WNE_P10: Average Wage 10 Years After Graduating From College
- ST_FIPS: State Codes
- ADM_RATE: Admission Rates (in percentages)
- SATVR25: Average SAT Verbal 25th Percentile Score
- SATVR75: Average SAT Verbal 75th Percentile Score
- SATMT25: Average SAT Math 25th Percentile Score
- SATMT75: Average SAT Math 75th Percentile Score
- SATWR25: Average SAT Writing 25th Percentile Score
- SATWR75: Average SAT Writing 25th Percentile Score
- completion: Completion Rates
- transfer: Transfer Rates
- LOCALE: Setting (City, Suburb, Town, or Rural)
- CCUGPROF: Carnegie Classification Undergraduate Profile (2 year vs 4 Year institution)
- Year: Year of Data set that Information Originally Came From

From **college.csv**, we made a data set called **scaled-colleges.csv**. All of the variables in scaled-colleges.csv are the same as those in **college.csv**, but we converted factors (qualitative variables) into dummy variables, centered the mean, and standardized all of the data. Since, we standardized the data, this makes our data more comparable because all of our variables now on comparable scales. This is really important because our β will be different depending on the scale that the variable is measured in. By centering and standardizing, we will not favor any coefficient. Ultimately, this will help us with our regression analysis so that we can find the most accurate model and have the best predictions possible.

4 Methods

4.1 Standardization

The colleges information data contains data measured in different scales such as admissions rate versus age (which are on very different scales). To prevent any biased weighting, we want to standardize the information before modeling.

First, we created inidicator variables for the qualitative variables discussed in the introduction. Next, we want to standardize the absolute quantitites to relative quantities (mean centering).

One reason to standardize variables is to have comparable scales. When you perform a regression analysis, the value of the computed coefficients will depend on the measurement scale of the associated predictors.

In addition to standardizing, we removed all observations that have NA's.

4.2 Regression Models (Technical Explanation)

Ordinary Least Squares Regression (OLS) If predictors (regressors) are correlated, the stability of the $\hat{beta}$ decreases, meaning, every estimate of β could be very different and not converge to the true population coefficient.

$$Balance = \beta_1 Married + \beta_2 Income + \beta_3 1stGeneration + \dots$$

In order to find the best explanatory variables, we use model selection and cross validation. Using AIC and BIc, we find which variables would be the most helpful in predicting the admission rate. Each of these methods gives us two models:

The model using AIC is,

ADM_RATE ~ UGDS + UGDS_BLACK + UGDS_2MOR + AGE_ENTRY + UGDS_WOMEN + FIRST_GEN + FAMINC + MN_EARN_WNE_P10 + ST_FIPS6 + ST_FIPS9 + ST_FIPS10 + ST_FIPS11 + ST_FIPS12 + ST_FIPS13 + ST_FIPS15 + ST_FIPS17 + ST_FIPS20 + ST_FIPS21 + ST_FIPS24 + ST_FIPS25 + ST_FIPS29 +ST_FIPS30 + ST_FIPS34 + ST_FIPS35 + ST_FIPS36 + ST_FIPS37 + ST_FIPS39 + ST_FIPS40 + ST_FIPS48 + ST_FIPS50 + ST_FIPS53 + ST_FIPS54 + ST_FIPS78 + SATMT25 + SATMT75 + SATWR25 + SATWR75 + SATVR25 + completion + transfer + LOCALE23 + LOCALE31 + LOCALE32 + LOCALE33 + CCUGPROF5 + CCUGPROF6 + CCUGPROF7 + CCUGPROF8 + CCUGPROF9 + CCUGPROF10 + CCUGPROF11 + CCUGPROF12 + CCUGPROF13 + CCUGPROF14 + CCUGPROF15 + Year2011

The model using BIC is,

ADM_RATE ~ UGDS_BLACK + UGDS_WOMEN + FAMINC + ST_FIPS6 + ST_FIPS10 + ST_FIPS12 + ST_FIPS20 + ST_FIPS21 + ST_FIPS24 + ST_FIPS25 + ST_FIPS30 + ST_FIPS34 + ST_FIPS35 + ST_FIPS36 + ST_FIPS37 + ST_FIPS50 + ST_FIPS53 + ST_FIPS54 + ST_FIPS78 + SATVR25 + SATMT25 + SATMT75 + SATWR75 + completion + transfer + CCUGPROF5 + CCUGPROF6 + CCUGPROF7 + CCUGPROF8 + CCUGPROF9 + CCUGPROF10 + CCUGPROF11 + CCUGPROF12 + CCUGPROF13 + CCUGPROF14 + CCUGPROF15

Using cross validation, we found that the best model out of these two is the one using AIC. The combination of these variables are the best at predicting admission rate.

Ridge Regression (RR)

RR is a variation of the minimization in OLS Regression but with a constraint of $||\beta||_2^2 < c^2$.

In vector form: $min\beta\ ||y - A\beta||_2^2 + \lambda||\beta||_2^2$

A difference in behavior of RR is that as λ increases, more weight is given to the second term in the minimization. This means that with a large λ, the β will be small.

The main advantage of RR is that it takes correlated parameters into account and does automatics parameter weighing.

Lasso Regression (LR)

LR is a variation of the minimization in OLS Regression but with a constraint of $||\beta||_1 < c$. With c, the constraint shape becomes a diamond and any pairs of β will likely contain zeros. Unlike RR, there is no explicit form of β.

In vector form: $min\beta\ ||y - A\beta||_2^2 + \lambda||\beta||_1$

The main advantage of LR is that it performs both parameter shrinkage through feature selection (sparsify regressors/predictors) and variable selection automatically.

Principal Component Regression (PCR)

PCR is based on principal component analysis. The goal of this method is to reduce the dimensions (created by the set of data points in n-dimensional space).

To do so, we want define a direction, the first principal component, that maximizes the the variability in the data set and set the second principal component perpendicular to this first principal component. As a result, each data point's coordinates will change to this new coordinate system.

Partial Least Squares Regression (PLSR)

PLSR, similar to PCR, is also a dimensionality reduction method. While PCR finds hyperplanes of maximum variance between the predictors and responses, PLSR projects the predicted variables and the observable variables into a new space.

The main advantage is that PLSR uses the annotated label to maximize inter-class variance. It takes into account of the classes and tries to reduce the dimension while maximizing the separation of classes.

4.3 Cross Validation and Train-Test Sets

Because we have limited amount of observations to build and test the model and we want to prevent bias, we will build and test the model using different subsets of the whole data set.

We built train sets using 75% of the observations and test sets of the remaining 25% by random sampling (without replacement). We repeated this process 10 times for a 10 fold cross validation when we ran the regressions.

5 Analysis

5.1 General

In order to complete cross validation, we first had to find the rows of information that contained complete entries, meaning that there is an input for every single column. We accomplished this by using the function complete.cases. Complete.cases creates a vector of true and false values, true when the row does not contain NAs and false when the row does have NAs. Each regression is then run on the parts of the train data that were complete.

Ordinary Least Squares Regression

As mentioned in the Methods section, we found the best model using model selection. The resulting model gave us the best predictive power in ordinary least squares. This model only includes a portion of the explanatory variables available, since model selection determines that some of the variables are not good predictors.

We build the model using the training data set and then using this model predicted the admissions rate using the testing data set. We then found the mean squared error (MSE) of these predictions. The MSE from OLS is 0.4476

Ridge Regression

When doing ridge regression, we started by looking at a ten-fold cross-validation. From cross validation, we were able to find the best model, which included finding our lambda or tuning variable.

The **MSE Plot of Ridge Regression** shows the relationship between MSE and log(lambda).

From running our cross-validation on the train data set, we find that lambda is 0.01. Using this lambda and regression, we predicted the admissions rate from the test set of data and found the MSE is 0.4389.

Lasso Regression

To run lasso regression, we utilized the glmnet package. In order to find the best lambda value for the regression, we created an array of 100 lambda values from 0.01 to 10^{10} and run a 10-fold cross validation on all these lambda values. We then proceeded to find the minimum lambda, 0.01.

The plot **MSE Plot Lasso Regression** shows the relationship between MSE and log(lambda).

Using this lambda and regression, we predicted the admissions rate from the test set of data and found the MSE is 0.4215.

Principal Component Regression

In order to find the best lambda value for this regression, we utilized *pcr()* with the attribute of "validation" as "CV". The resulting best number of components, lambda, is 98.

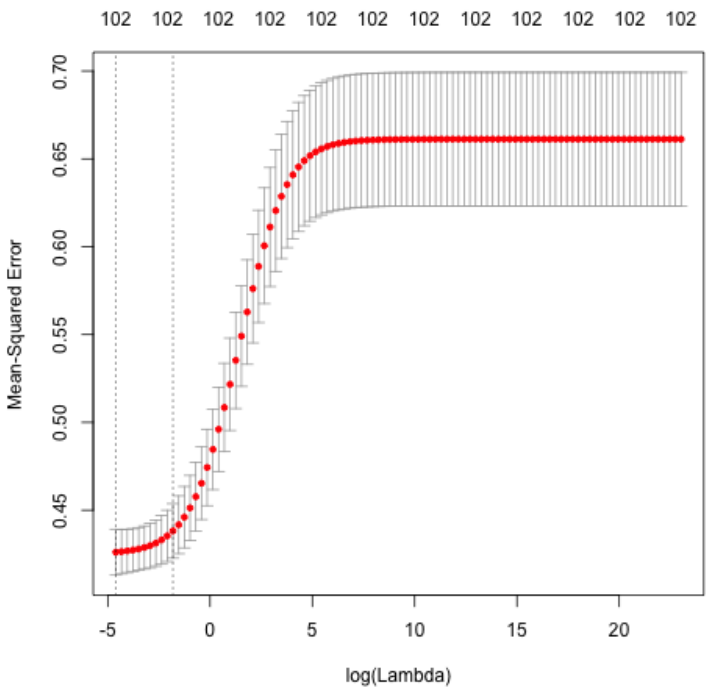

Figure 1: MSE Plot of Ridge Regression

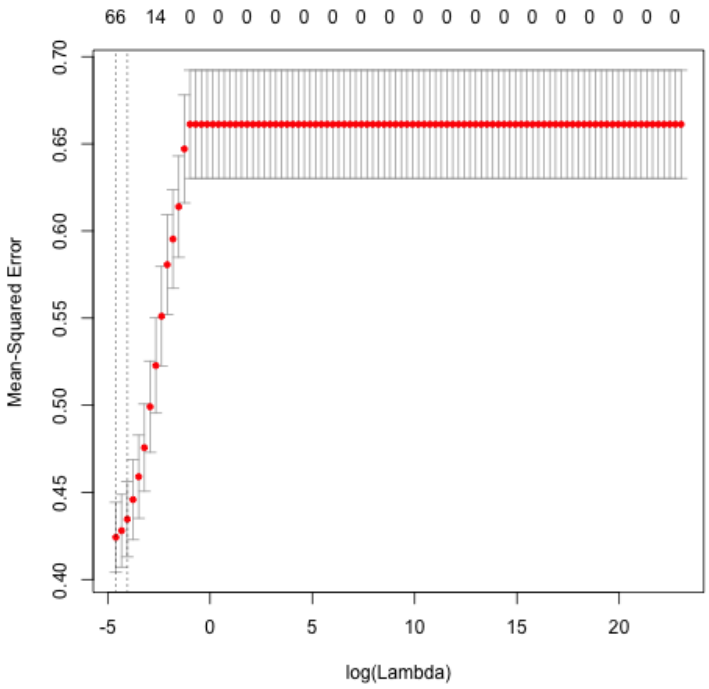

Figure 2: MSE Plot of Lasso Regression

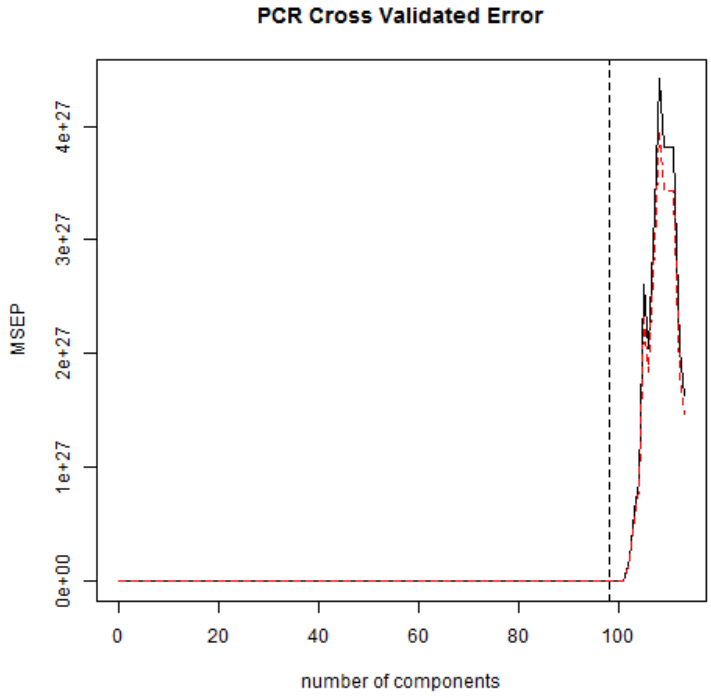

Figure 3: MSE Plot of Principal Components Regression

The plot **MSE Plot of Principal Components Regression** shows the relationship between MSEP and the number of components.

Using this lambda and regression, we predicted the admissions rate from the test set of data and found the MSE is 0.4455.

Partial Least Squares Regression

In order to find the best lambda value for this regression, we utilized *plsr()* with the attribute of "validation" as "CV". The resulting best lambda is 51.

The plot **MSE Plot of Partial Least Squares Regression** shows the relationship between MSEP and lambda, the number of components.

Using this lambda and regression, we predicted the admissions rate from the test set of data and found the MSE is 0.4509.

6 Results

In order to figure out which is the best model, we decided to compare the mean squared error (MSE) values and pick the model and method that produced the smallest MSE value. The smaller the MSE, the higher the prediction accuracy of the model. When Looking at the table Test MSE Values for the Regression

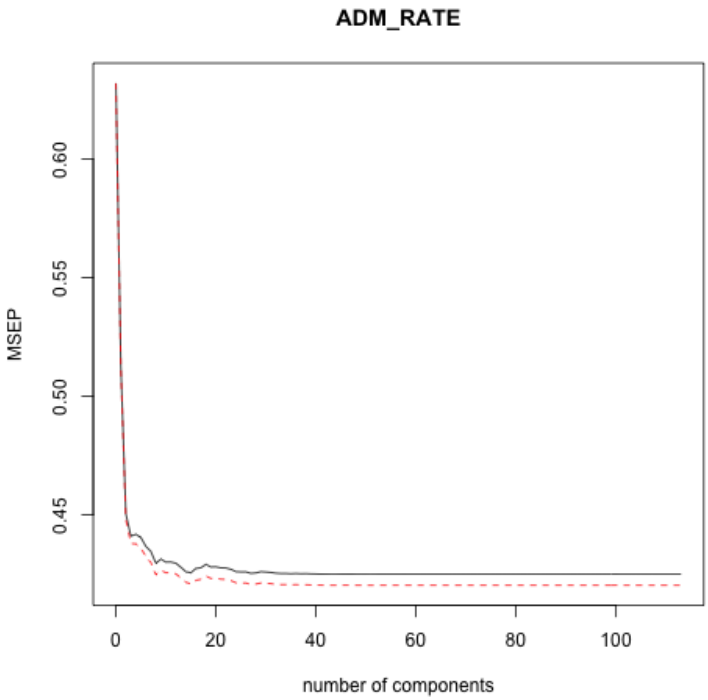

Figure 4: MSE Plot of Partial Least Squares Regression

Techniques, we found that all the MSEs were very similar with the range being 0.0294. Of all the regression methods, we found that the one that produced the smallest MSE is lasso regression, where MSE = 0.4215. So, we will use the equation produced via lasso regression in our app in order to help predict admission rates.

However, one issue that we ran into when using lasso to predict the coefficients for the full model is that it turns almost all of the variables into 0. This reduces the model to a state where it is not reliable and does not make that much sense to us. Since it is likely that lasso regression will turn all of our variables into 0s, we do not want to use this model. Instead, decided to solve the reliability issue in our app by calculating a model using a smaller subset of schools and ridge regression to find the intputs. When using our app, we will first input the school's OPEID in order to get information about the school. Focusing in on the admission rate, our app will randomly choose 20 schools with a smaller admission rate and build a regression model using lasso regression. From this equation, we will predict what our client's new admission rate is using the scaled data for our client's school as the inputs. The output of our model will be an admission rate in terms of scaled data, but in our app, we have a formula that will convert that scaled admission rate into an admission rate that is understandable and comparable with their current admission rate. Additionally, we can select the variables that we want to input and using various variables and various amounts of variables, that will change what our prediction is.

To better illustrate how this works, we will walk through an example together. Let's pretend our client

Table 1: Test MSE Values for the Regression Techniques

Regression	MSE
OLS	0.4476179
Ridge	0.4388746
Lasso	0.4214757
PCR	0.4455302
PLSR	0.4509218

is Alabama A&M University. We input some of attributes into our app to find some schools that are similar to Alabama A&M and found schools that are similar but have a lower admission rate. Alabama A&M's admission rate is .56, and through our app, some similar schools with lower admission rates are Villanova University, Chicago State University, Franklin and Marshall and etc. Using the regression model that was creating using the sample of schools that are similar but with a lower admission rate, our model predicts that if Alabama A&M becomes closer to those other schools in terms of size, percentage of women, percentage of black students, percentage of white students, and the average age when entering college, then it's new admission rate could be .54. This means that the admission rate is lowered by .02, which is pretty reasonable for changing admission rates within a year.

We also tried different combination of variables. For example, when we only used inputs such as age of entry and size, it predicted that our admission rate would be .57, which is actually higher then what our original admission rate was. However, when we used size, percentage of women, percentage of black students, percentage of white students, percentage of Hispanic students, and the average age when entering college, we were able to lower the admission rate to be .52, which is much lower that what we started off with and even lower then the prediction that we previously found. Being able to play with these variables and inputs is important because it helps us to identify the areas and profiles in which there needs to be changes for each school and we can help them to improve their rates by various percentages with ability to focus in on certain areas.

7 Conclusion

Through the use of linear methods, we were able to find the method that would produce the best possible model for our data. This method turned out to be lasso regression. While lasso regression increases bias, it decreases variability. With a smaller variance, the predictions are closer to the true model meaning that this may be a better model.

Having a good linear model is very important to our project because our entire shiny application and consulting recommendations depends on it. From our linear model, were are able to predict what the admission rates would be based on a sample of schools that are similar but have a smaller admission rate.

Additionally our shiny application will allow us to make the best recommendations to our client because it tells us exactly how much their rate could be lowered by and which areas the can improve on in order to get their rate to what we predicted.