Part IV - CollegeScorecard

Jacob Waldor & Moe Sunami

Introduction

College ScoreCard[1] provides institution-level data for approximately all institutions of higher education in the United States, including college type, costs, graduation & retention, financial aid and debt, test scores, earnings, and demographics. To be included in the dataset, institutions must fulfill essential functions of a college to make them eligible for federal student financial assistance from Title IV programs. Colleges that are not in the dataset may not offer the equivalent of 1 weeks of education, or may not provide degrees towards gainful employment or the liberal arts.

For this project, we seek explain the prestige of U.S. institutions of higher education with four-year undergraduate programs, using admission rate as a measure of prestige. We do not have a random sample; we begin with a dataset that contains approximately entire population, and analyze the observations among them that have complete data. So, our goal is to describe the relationships that are present in our data, as opposed to evaluating relationships that can be generalized to a larger population.

- Which set of regressors most succinctly summarizes the variance in admission rate? Admission rate is a mysterious quantity, so we want to demystify it by seeing if we can find a small set of variables that explains the variance in it well.
- What theories can explain variance in admission rates? We might expect that, if a college is more selective, its graduates earn more after graduation.

Variables of Interest

Of the 2996 variables in our dataset, we have selected 15 which seemed to have reasonable theoretical relationships to admission rates and were available from recent years (Table 1). Additional feature engineered variables are indicated with a *.

Note from Part II that we have transformed the `ENDOWBEGIN` variable with a log function to meet LINE conditions.

Note on Missing Values

Although we have identified 1465 colleges in CollegeScorecard offering four-year undergraduate programs, only 1085 of them (approximately three-fourths) have complete data for all of our relevant predictors. We suspect that those 1085 colleges are not a random sample, and that there is a pattern to which colleges have complete data. For example, none of the Claremont Colleges have complete data because its 5-year loan repayment rate is noted as "PrivacySuppressed," presumably because of small class sizes. Hence the p-values that may accompany hypothesis testing in this report, or confidence and prediction intervals, must be taken with a grain of salt because it is *not* the probability of observing more unusual data, given H_0 were true, if we were to have another "random" set of observations.

[1] This data is published by the Office of Planning, Evaluation, and Policy Development at the U.S. Department of Education and is made available at collegescorecard.ed.gov. The data is drawn from data reported to IPEDS (Integrated Postsecondary Education Data System) from parent institutions.

Table 1: Relevant variables

Name	Description
HBCU	Flag for Historically Black College and University
WOMENONLY	Flag for women-only college
ADM_RATE_ALL	Admission rate for all campuses rolled up to the 6-digit OPE ID
SATMT25	25th percentile of SAT scores at the institution (math)
SATMT75	75th percentile of SAT scores at the institution (math)
SATVRMID	Midpoint of SAT scores at the institution (critical reading)
SATMTMID	Midpoint of SAT scores at the institution (math)
SAT_AVG	Average SAT equivalent score of students admitted
UGDS	Enrollment of undergraduate certificate/degree-seeking students
PCTPELL	Percentage of undergraduates who receive a Pell Grant
PCTFLOAN	Percent of all undergraduate students receiving a federal student loan
COMPL_RPY_5YR_RT	Five-year repayment rate for completers
MD_EARN_WNE_P10	Median earnings of students working and not enrolled 10 years after entry
ENDOWBEGIN	Value of school's endowment at the beginning of the fiscal year
GT_THRESHOLD_P10	Share of students earning more than a high school graduate (threshold earnings) 10 years after entry
*LOWSAT	Equal to SAT_AVG when SAT_AVG $<$ 1300, 0 otherwise
*HIGHSAT	Equal to SAT_AVG when SAT_AVG $>=$ 0, 0 otherwise
*SATmathinterquartile	SAT math interquartile range (25th to 75th percentile), indicating spread
*SATverbalmathdiff	Difference between median verbal and math scores on the SAT, indicating level of focus on either humanities or STEM

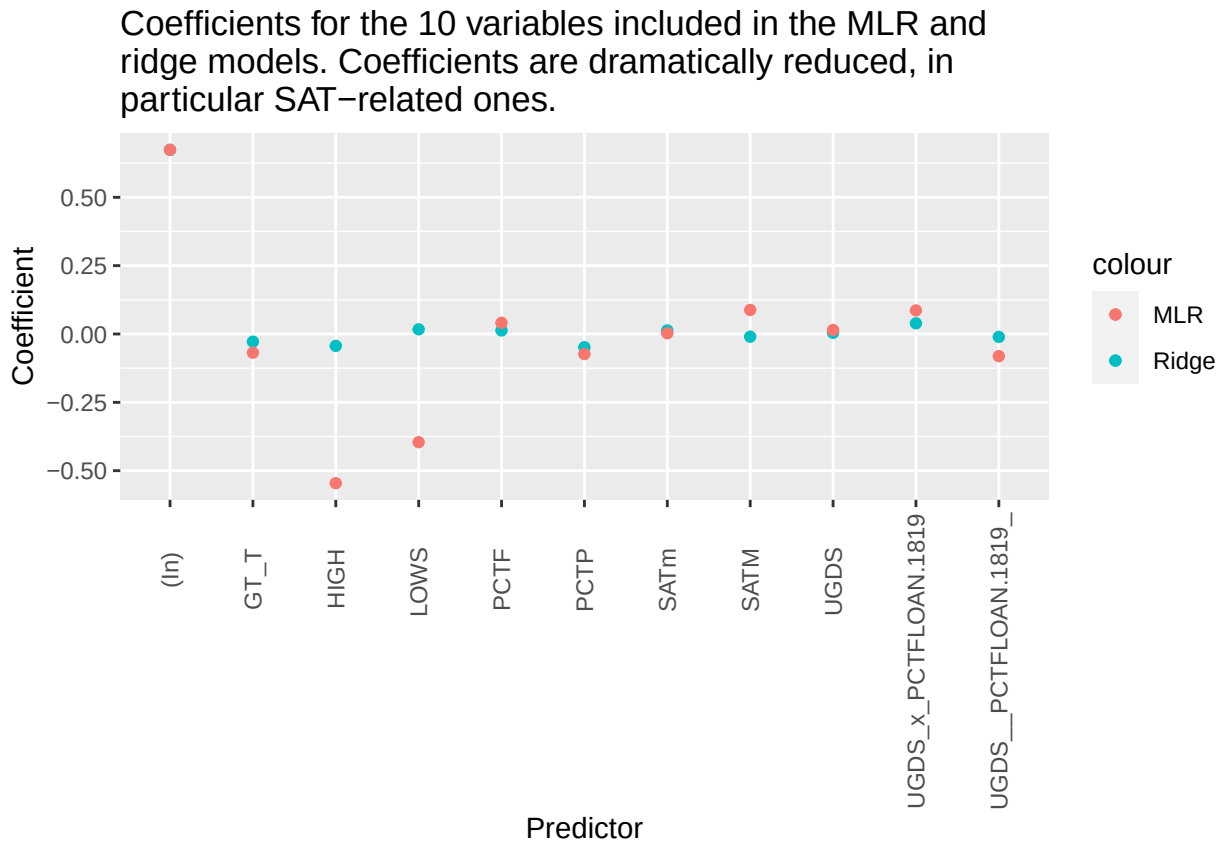
Section 1

Ridge Regression

To prepare our data for ridge regression, PCR, and LASSO, we normalize the dataset by finding the mean and standard deviation of each variable and replacing each value with the number of standard deviations by which the value differs from the variable mean. We also re-generate our MLR model using normalized data so that the coefficients are comparable to the coefficients in the other models.

We run a ridge regression on the set of all ~30 predictors that were fed into the forward selection model.

Below is a plot of the ridge coefficients vs. the the MLR coefficients. Below, we only examine the coefficients for the 10 variables that were included in the MLR model—the other ~20 ridge coefficients are excluded. Since we normalized the dataset, we can compare the magnitudes of these coefficients as well as the directions.



The leftmost dot in the figure above corresponds to the intercept. The intercept is exactly the same for both models—a red dot is visible because it overlaps with the blue dot.

We see that ridge regression produces dramatically different coefficients for HIGHSAT and LOWSAT. Both of these coefficients were very large in the original model. We also notice that some coefficients, including UGDS, switch signs. This is reasonable given that UGDS has high p-value.

Lastly, we turn our attention to variables that were not included in the forward model. We compare the average magnitude of coefficients for all variables in the ridge regression to the average magnitude of the coefficients for ridge regression variables that also appeared in our MLR model. The average magnitude for the former is 0.04, whereas the average magnitude for the latter is 0.08. This difference is natural; the MLR model includes the variables with relatively high explanatory power, so these variables should have higher coefficients in the ridge regression.

LASSO

We now turn our attention to the LASSO.

Unlike ridge regression, the LASSO zeroes out some of the coefficients. Therefore, we can think of the LASSO as comparable to the forward selection algorithm that generated our MLR model. To compare the LASSO to the forward selection algorithm, we look at which coefficients fall out when we run the LASSO model and compare it to which variables made it into our MLR model. Unsurprisingly, a bunch of interaction terms fall out of the model. Also removed are MD_EARN_WNE_P10.1819, WOMENONLY, SATverbalmathdiff, SATMT25, and SATMTMID, which is unsurprising since they also were not included in our MLR model.

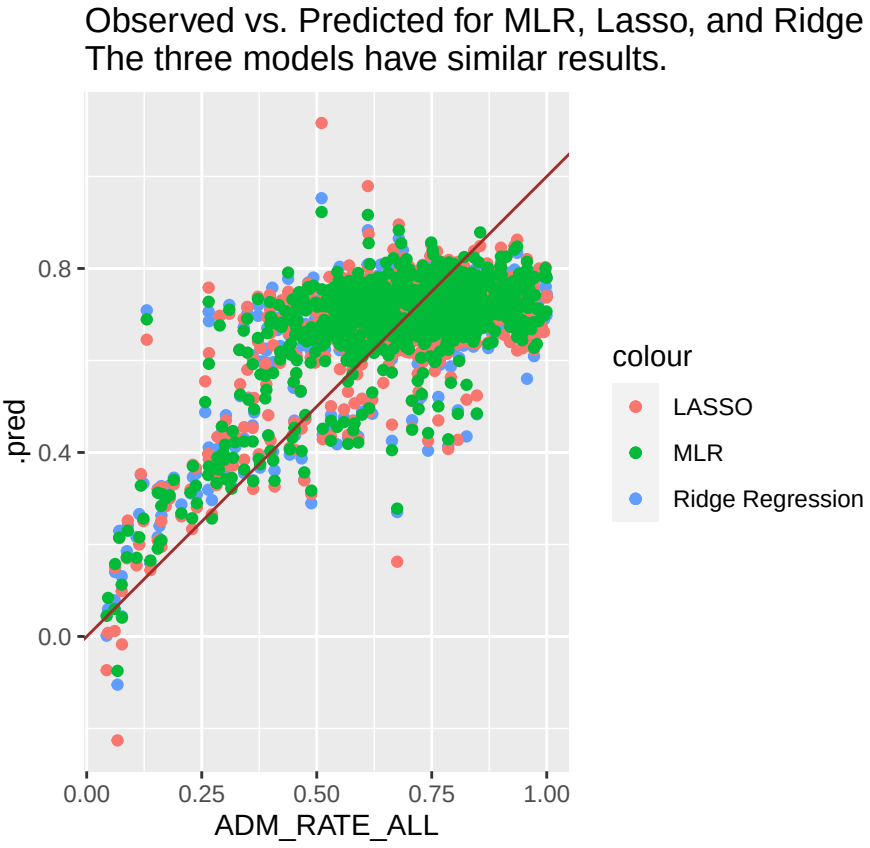
While it did not fall out of the MLR model, UGDS fell out of this one. Two interaction terms involving UGDS were included, however, whereas only one was included in the MLR model. The one interaction term in the MLR model that has UGDS is also included in this model.

Our LASSO included some variables that did not end up in our MLR model, including GT_THRESHOLD_P10.1819, MD_EARN_WNE_P10.1819, PCTFLOAN.1819, and SATVRMID. The first three did show in the MLR's interaction terms though. Although SATVRMID did not appear in any of the MLR model's interaction terms, its coefficient in the LASSO model is pretty low.

Overall, it looks like LASSO selected variables that were not much different from our MLR.

Outcomes vs. Predictions for all three models

We generate a graph with outcomes vs. predictions for all three models together.



First, we notice that the three models have very similar results. This is surprising given how ridge regression seems to reduce some of the the coefficients so much. It could be that ridge spreads the explanatory power

between collinear variables more, leading to similar results but with more variables taking an explanatory role in the model.

Little is apparent other than this. One strange thing is that LASSO seems to produce the most points that are very far away from the line.

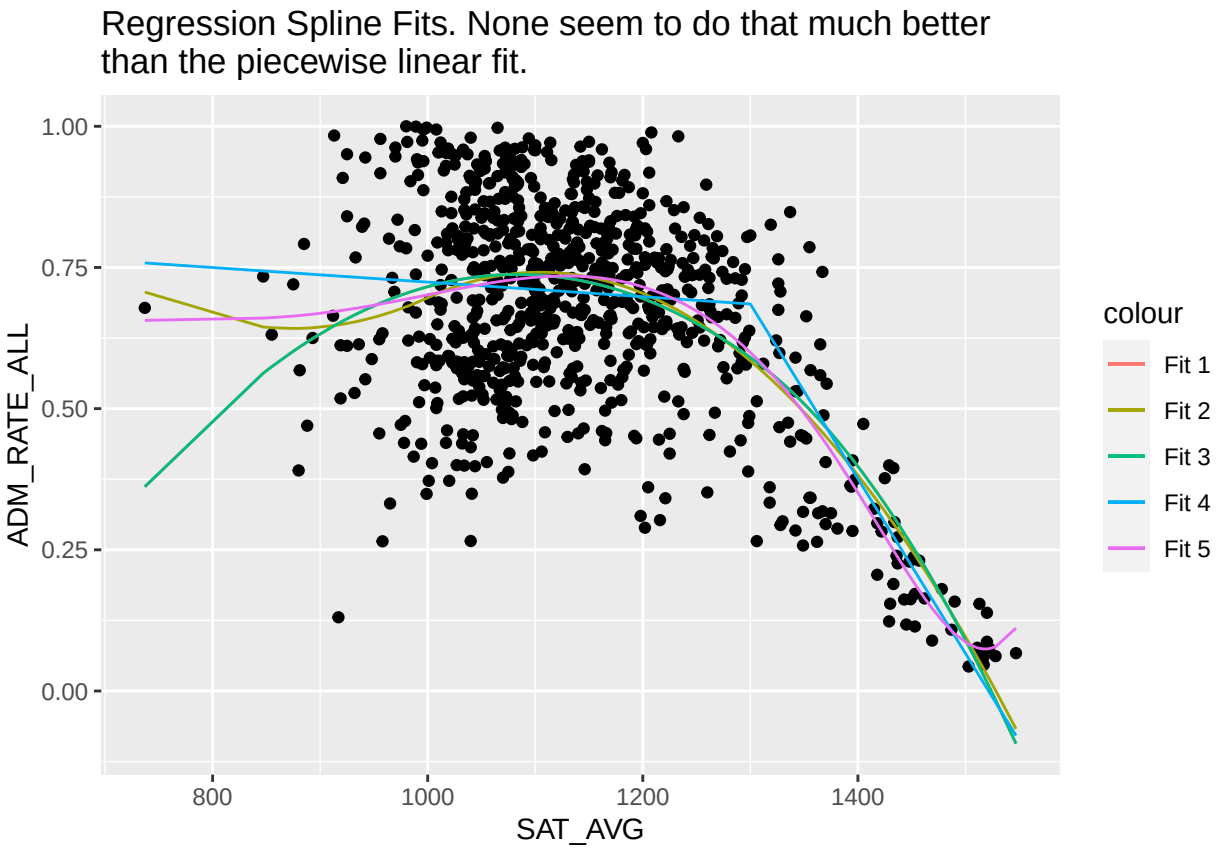
Lastly, it is fascinating how the shape of this plot is similar to the `SAT_AVG` vs. admission rate plot. This similarity suggests that `SAT_AVG`—and variables collinear to it—could really possess the bulk of the explanatory power. The plots also reveal that we are much better at predicting the admission rates of selective schools.

Spline & Kernel Smoothing for SAT_AVG

Because we witnessed some interesting relationships between `SAT_AVG` and `ADM_RATE_ALL` in part 3, we will run splines with `SAT_AVG` and `ADM_RATE_ALL` here.

We run splines with the following knots:

1. At SAT_AVG = 1300 with degree 2
2. At SAT_AVG = 1000, 1300 with degree 2
3. At SAT_AVG = 1300, 1600 with degree 2
4. At SAT_AVG = 1300 with degree 1
5. At SAT_AVG = 1300 with degree 5

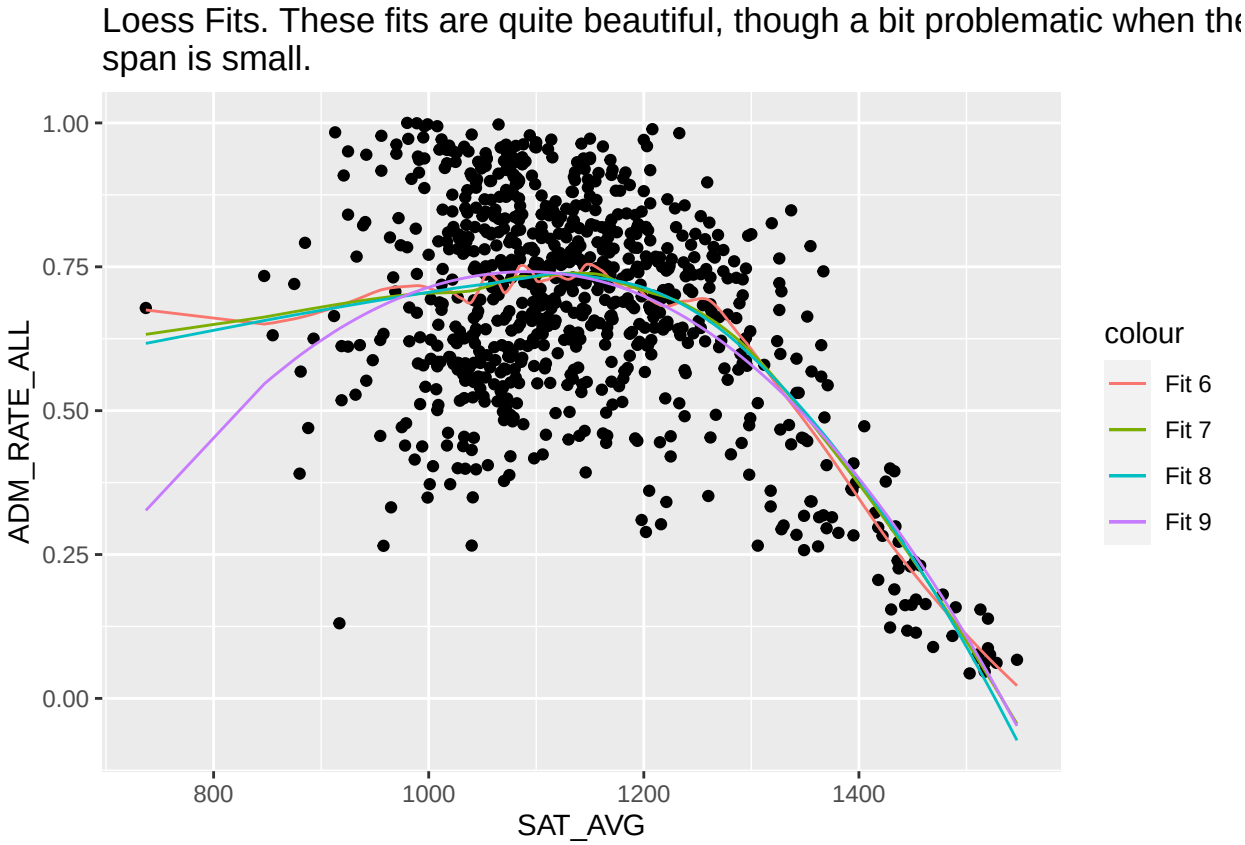


We see that the purple curve (Fit 5), which has a knot of degree 5, has extra turning points that are not

appropriate for the data. The extra knot at 1000 for the dark green curve (Fit 2) allows it to fit the data more closely. Overall, it is not clear that any of the splines do much better than the blue curve (Fit 4), though we haven't rigorously assessed this using DOF-adjusted metrics.

Now, we run loess with the following spans:

1. 0.2
2. 0.5
3. 0.8
4. 2



The loess models all do a good job of capturing the changing slope in the data. Given how loess works, this makes sense. The span associated with the red curve is too small, leading to erratic behavior, but the erratic behavior isn't that dramatic

Overall, I'd choose the blue spline model (Fit 4) to make future predictions. The blue spline model is very interpretable—we have a certain slope up to 1300 and a different slope after. It succinctly captures my intuition about what is happening with the slope of the points. We have a relatively flat slope up until about 1300, and then a greater slope later. However, this model is not differentiable, and so it is unable to capture how the slope might be changing continuously in a neighborhood around 1300. The ideal model might be linear in the beginning, then become curved, and finally become linear again.

Conclusion for Section 1

Our findings here were more or less as expected. Adding degrees of freedom to the spline models (through more knots, higher degree) did not seem to dramatically help capture the relationships in the data. Although

we chose the blue spline model for its simplicity, loess did a great job. Impressively, it automatically captures relevant features. For example, we did not need to specify the locations of knots to get it to work. How cool! It would be great if we had ways of more easily interpreting loess models or performing ablation to remove unnecessary degrees of freedom from it.

Trying many models here leaves us with the question of what's really going on with `SAT_AVG` and `ADM_RATE_ALL`'s relationship. It would be fascinating to try to use these data to find a model that both fits the data well and has some form of explanation. A piecewise linear relationship (as in the blue spline model) seems too simplistic. The real world tends to behave in a more differentiable way. Ideally, we want to see why some schools exhibit the linear relationship and others don't. We could take many random subsets of schools with SAT scores around 1300 (e.g. 1000 to 1500) and see which subsets exhibit the most linearity. Then, we could use a classifier to see what distinguishes those subsets from the others.

Section 2

Principal Components Analysis

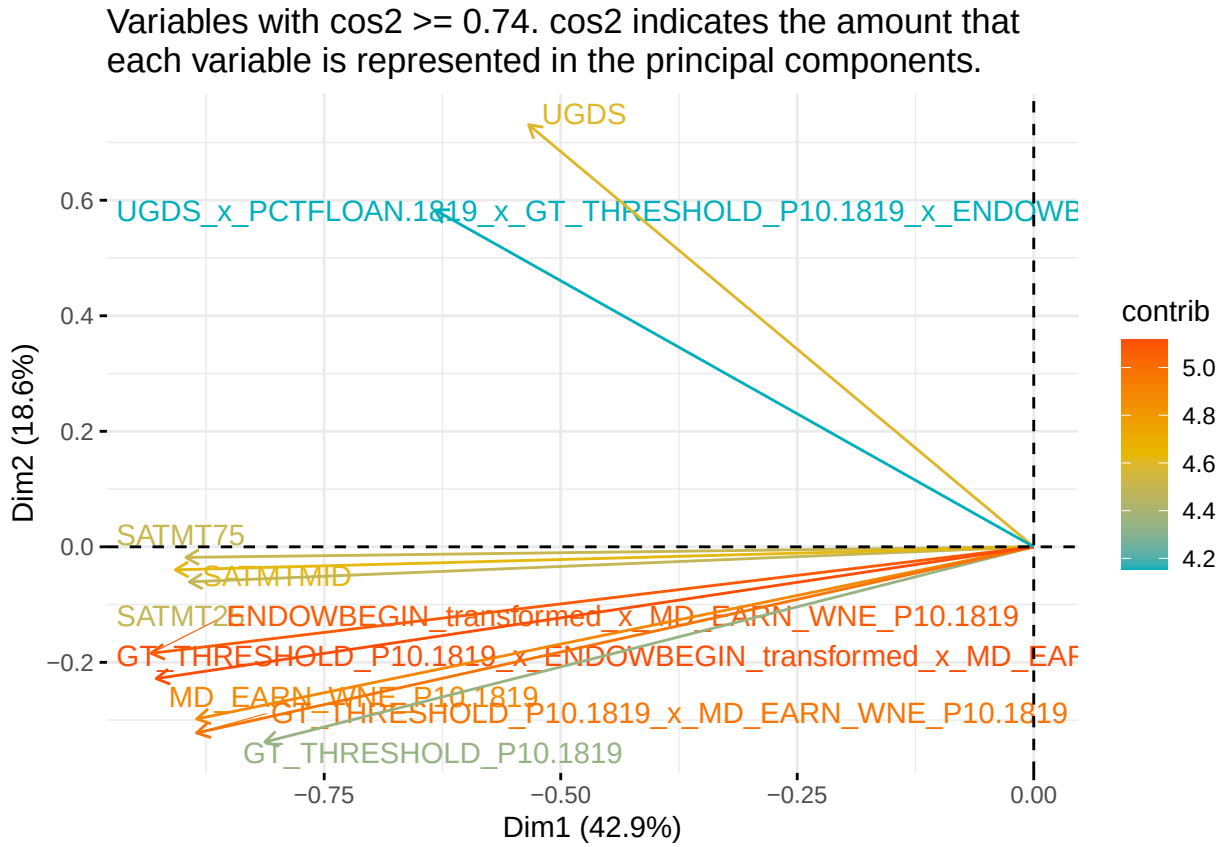
We apply Principal Components Analysis (PCA) to the explanatory variables and then perform a Principal Components Regression using the principal components. PCA compresses the features of all observation into p features that capture the key ("principal' ') axes of variation for the observations. We can visualize PCA as drawing a straight line in hyperspace such that when we project the positions of the observations to that line, the variance of the positions of the projections is maximized. We continue to draw such lines—with the additional requirement that each line is orthogonal to the lines that have already been drawn—until we have p lines. The p lines represent our p principal components. After generating the principal components, we can use them as features in a linear model, giving us Principal Components Regression.

PCA helps us get a more interpretable model for explaining the variance in admission rate. Multicollinearity in part III made it difficult to interpret coefficients because the predictors are correlated with one another (e.g. because `HIGHSAT` and `LOWSAT` have a Pearson correlation coefficient of -0.97, we cannot interpret each coefficient to mean "the change in response when other variables are held constant.") By understanding the key axes of variance for the predictors we are using, we can eliminate multicollinearity and understand which factors independently explain admission rate.

Principal Components Regression involves some assumptions. Since it is essentially an MLR model—it just involves variables that have been transformed into principal components—it assumes that our principal components satisfy the LINE conditions. However, provided that our original set of variables satisfy the LINE conditions, our principal components as well as the target variable will also satisfy the LINE conditions. Note that the previous statement only holds when we include all principal components in our regression. If we omit some, we can lose linearity.

Although our original dataset does not satisfy equal and normal variance perfectly, for our purposes, we don't really need to worry about this—we are not trying to generate a predictive model, but rather to see which factors can explain the variance in admission rate.

We visualize our principal components as follows:



We analyze the first principal component using the figure above. Variables involving SAT and `GT_THRESHOLD` (which indicates percentage of graduates earning above an average high school graduate) tend to live in the first principal component. Since earnings and SAT are generally related to intelligence, both of these variables are likely strongly related to intelligence. Therefore, the first principal component could relate largely to intelligence of students. Finally, note that the figure suggests that SAT has a large role in the variance of the explanatory variables. This is interesting given our hypothesis that SAT has a large role in explaining variance in admission rate.

The second principal component is more closely connected to UGDS. It could reflect the name recognition of the institution, since larger institutions have greater name recognition.

Lastly, we notice that interactions that involve similar variables are close together, which makes sense. For example, UGDS and an interaction term involving UGDS are close together.

Below, we provide the results of our Principal Components Regression:

```
## Data:    X dimension: 813 29
##  Y dimension: 813 1
## Fit method: svdpc
## Number of components considered: 29
## TRAINING: % variance explained
##               1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X               42.92    61.48    75.52    81.36    85.13    88.52    91.31
## ADM_RATE_ALL    16.94    16.94    26.85    39.02    39.11    39.23    39.37
##               8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X               93.80    95.79     97.00     98.14     98.93     99.40
## ADM_RATE_ALL    39.43    39.94     41.48     41.48     42.75     42.76
##               14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
```

```
## X                99.63     99.80     99.93     99.96     99.97     99.99
## ADM_RATE_ALL     43.69     44.43     44.48     44.72     44.86     44.87
##               20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## X                99.99    100.00    100.00    100.00    100.00    100.00
## ADM_RATE_ALL     45.51     47.55     47.85     47.85     47.92     47.99
##               26 comps  27 comps  28 comps  29 comps
## X               100.00    100.00    100.00    100.00
## ADM_RATE_ALL     48.01     48.01     48.23     48.53
```

We notice that the first 10 variables in the model take care of 41 percent of the variance. On the other hand, in our MLR model, the first 10 varaiables take care of 46% of the variance. This makes sense given how whereas our MLR model was created by forward selection, which chose variables according to what helped explain variance in Y, principal components were not chosen with Y in mind.

In fact, we see that 39% of variance is explained by just the first 4 components, whereas having all 29 only explains 49% of the variance. The first component explains 16.94% of the variance—more than any other variable—revealing that the variables above that we considered to be intelligence-related are very powerful Although they are not associated with the variance in the explanatory variables as much as the first component, the third and fourth components explain over 10% of the variance in the target variable, making them more impressive than the first component as predictors. In contrast, the second component does virtually nothing to explain outcome variance. This is really cool—there could be a deep explanation for this.

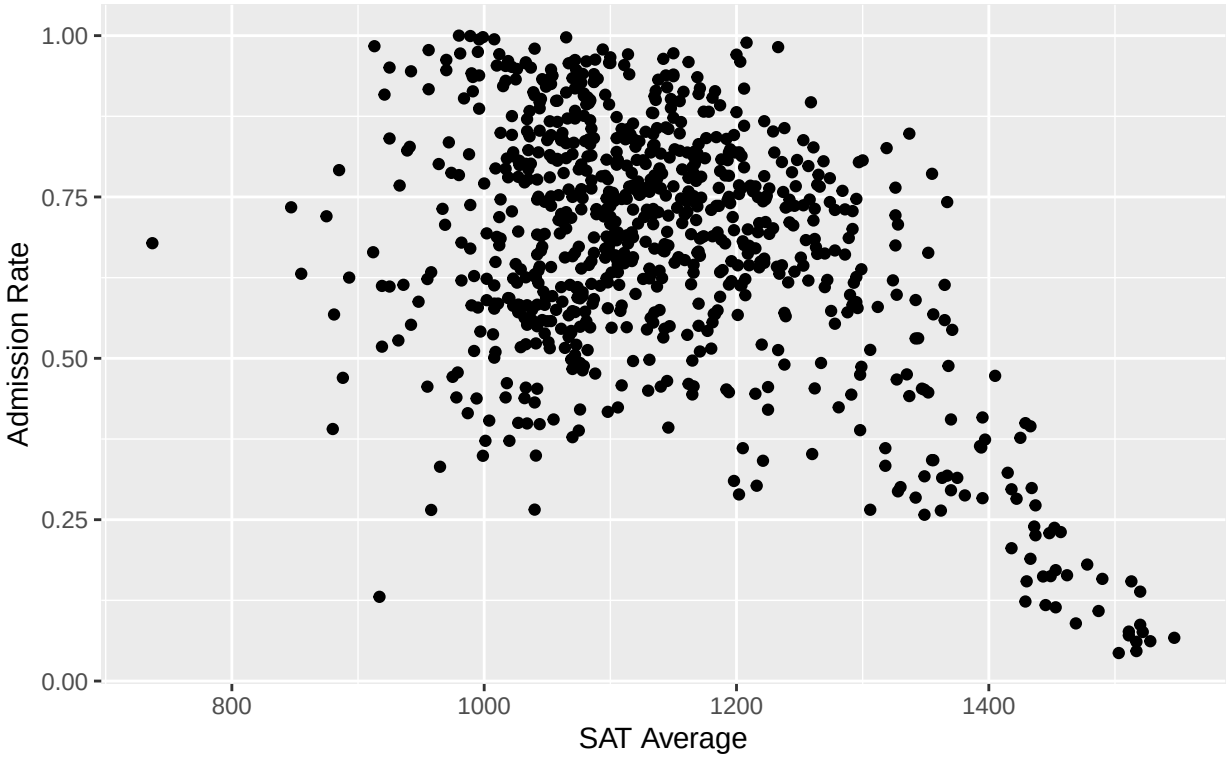
Like the first component, the third and fourth components seem to involve a lot of test-score related components. Interestingly, the fourth component has relatively large coefficients on interquartile ranges of test scores as well as endowment. Since interquartile ranges reflect STEM focus, the fourth component could relate to the financial power of an institution.

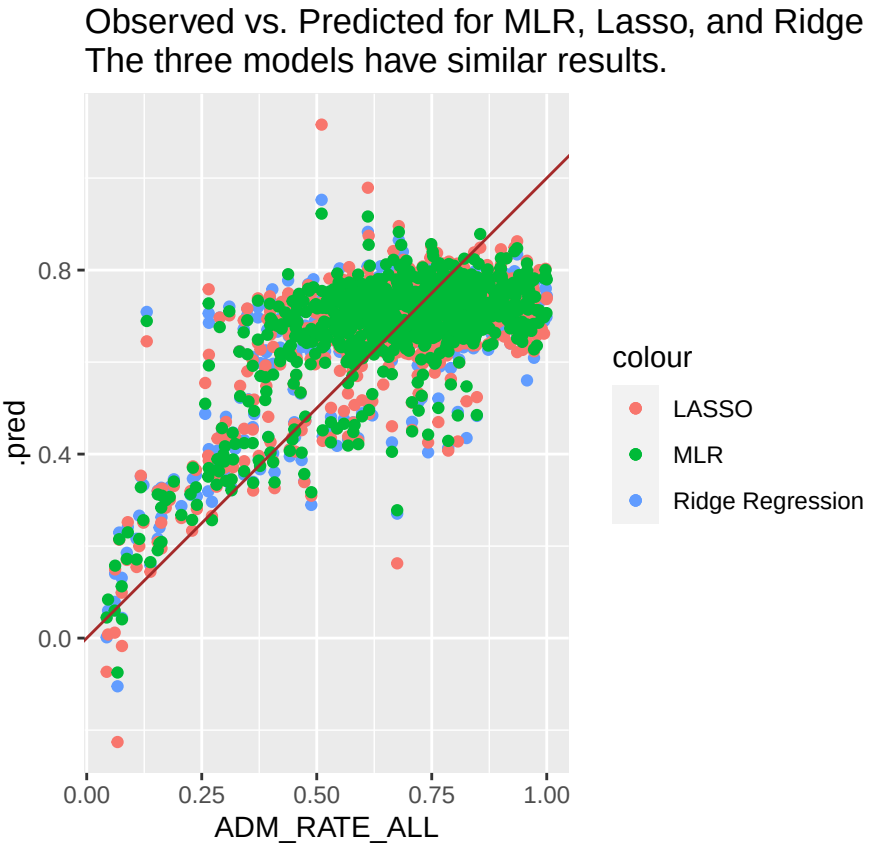
Section 3

CollegeScoreCard Project Summary

Our most interesting finding is the relationship between SAT scores and admission rates among undergraduate, 4-year schools receiving Title IV funding. Our MLR, ridge, and LASSO models did not yield clear insights about which variables are related to admission rates. At the same time, coefficients on SAT-related variables in these models as well as the principal components regression suggest that SAT score has the most powerful role in these models. In addition, the prediction vs. outcome plot for our MLR model (as well as ridge and LASSO) resemble the plot of `SAT_AVG` vs. `ADM_RATE_ALL`, suggesting that SAT alone is responsible for much of the variance that our MLR model explains. Further supporting this, doing single variate regression with `SAT_AVG` and `ADM_RATE_ALL` yields an R^2 of 0.22, compared to an R^2 of 0.4-0.5 in our more complex models.

Admission rate and SAT Average. There is a stark change in relationship when SAT is about 1300.





The relationship between `SAT_AVG` and `ADM_RATE_ALL` is fascinating because the relationship changes in such an obvious way and so rapidly. Above an SAT score of 1300, the relationship immediately changes from a nearly uncorrelated one to a strongly correlated one. So, one of the most appropriate avenues forward for shedding light on admission rate would be to see why some schools exhibit the linear relationship and others don't. We could take many random subsets of schools with SAT scores around 1300 (e.g. 1000 to 1500) and see which subsets exhibit the most linearity. Then, we could use a classifier to see what distinguishes those subsets from the others. It could be that some variables in the dataset that we haven't considered will help with the classifier, such as the geographic locations of schools. It could be that the socioeconomic status of students is the main distinction between the two categories, so we could seek more data on that, such as attendance of private schools.

On the other hand, having SAT in the model could have prevented us from getting insights into the relationships between other variables and admission rate. It is natural that scores on the SAT, an admissions test, should be related to admission rate. To see whether we might have some surprising variables related to admission rate, it is also important to generate a model that omits SAT.

In future data analysis, it could be useful to focus more on single-variate regression than multivariate regression. Since our goal is not to generate a predictive model but rather to understand the relationships between different variables and admission rate, using single variate regression will help avoid the ambiguity that MLR produces in how each variable is individually related to the target.

Future analysis would also benefit from a more clear delineation of what our population is. Although we have clearly defined our population as four-year undergraduate institutions receiving Title IV funding, it is not clear what the significance of Title IV funding is. Judging by the non-constant relationship between SAT scores and admission rate within our population, it seems that some of the schools do not match our Pomona student intuition of what a college is—we would expect SAT scores to always have a negative relationship with admission rate. We could think of this situation as involving perhaps two populations—the schools whose admission rate is relatively unrelated to SAT score and the schools whose admission rate is related.

The classifier mentioned above could help us more clearly define what these two populations are.

To conclude, we found some unexpected insights from our college data. We very well might be better off conceiving of our dataset as containing two populations rather than just one. Our project serves as a reminder of the importrance of specifying a population according to which research questions we want to bring to data. In our case, we were initially interested in what explained the admission rates of prestigious colleges we were interested in, but then found that the dataset consisted of many colleges that were much different. In response, we could have formed a rigorous computational definition of top colleges to more precisely specify our population, or we could have changed our research question to classifying the colleges in the dataset into meaningful sub-populations.