Analyzing university admissions outcomes after the implementation of the Top 10% law in Texas

Ben Inbar, Cliff Lee, Daria Dubovskaia, David Simbandumwe, Jeff Parks

DATA 621 Business Analytics and Data Mining
Group 1 - Final Project
May 21, 2023

Table Of Contents

Abstract

This study investigates the impact of the 1996 *Hopwood* case and subsequent Automatic Admissions Programs (AAPs) on the racial and ethnic composition of admissions at Texas A&M University. We estimate the extent to which these universities used affirmative action before the ban, and we examine how admissions officers at these universities adjusted the relative weights given to key applicant criteria after the ban and AAPs went into effect. We model the extent to which these new regulations succeeded in preserving minority admission rates at pre-*Hopwood* levels after examining whether changes in relative weights favored minority applicants. We discovered that most colleges followed the *Hopwood* rule, so that direct advantages offered to black and Hispanic candidates vanished (and, in some circumstances, became disadvantages). While some universities made adjustments in applicant characteristics that benefited underrepresented minorities, overall, the share of admitted students who were black or Hispanic did not remain constant despite the implementation of AAPs. These findings suggest that the new policies were not a reliable substitute for considering race and ethnicity in admissions decisions.

Keywords

- Hopwood, university, admission, diversity, ethnicity, regression

Introduction

Disparities in access to higher education, particularly for students from underrepresented groups, remains a national issue. Affirmative Action policies in university admissions, long considered a solution by some, faced resistance in the 1990's and were eventually outlawed in Texas in 1996, leading to a sharp decrease in minority enrollments in Texas universities.

In 1992, a legal challenge was brought by prospective law student Cheryl Hopwood against the University of Texas School of Law over the use of race-based admissions criteria. The resulting 1996 *Hopwood* decision handed down by the Fifth Circuit Court of Appeals declared the only justification for affirmative action as "rectifying present effects of past discrimination," and banned the direct consideration of race or ethnicity in the admissions criteria for public universities.

This decision extended to scholarships, financial aid, and recruitment, leading to a significant decline in minority student representation for the 1997 class and beyond in Texas' public institutions. Over the following decade several other states followed suit, banning affirmative action in public university admissions and causing concerns about maintaining minority enrollment levels.

In an effort to prevent further decline, the Texas legislature enacted H.B. 588, also known as the Top 10% law, granting state high school seniors graduating in the top decile of their class automatic admission to any public university. This effort to boost enrollment by students from traditionally under-represented populations and districts - without specific reference to race or ethnicity - was fully implemented by the fall 1998.

In this analysis, we will examine the effects of the Affirmative Action ban and the *Hopwood* decision on the makeup of the applicant cohorts at Texas A&M University, to understand the resulting impacts on the diversity of incoming students.

Literature review

Several studies we reviewed used SAT test score transmission data as a proxy for the quality and types of colleges students ultimately applied to. These data were used to examine the potential impacts of affirmative action bans and Automatic Admissions Programs (AAPs) from a number of angles.

Card & Krueger (2005) looked at test transmission data to investigate whether "highly qualified" (based on SAT score or High School GPA) minority students' application behavior was affected by the bans. While acknowledging that overall admissions rates for Black and Hispanic students had fallen 30-50% at affected public universities, these particular students' score-sending behavior (and presumably their expected admissions probabilities) did not change significantly. This was an interesting finding, but we felt this study failed to address the obvious implications for the majority of students who did **not** meet this score-based criterion, which itself is known to often correlate to minority status via economic and societal factors.

Long (2004) examined similar data but with a broader mandate to understand the effects on **all** minority students' score-sending behavior. Using simulations based on count models, Long's study predicted a decrease in reports sent by minority students and an increase in non-minority score reports sent to selective programs. This suggested changes in student expectations of acceptance and potential decreases in minority enrollment in selective programs.

Overall, the studies based on SAT score transmission data generally acknowledged the disconnect between application-related behavior and actual admissions, and the need for complex estimates of correlation.

Long & Tienda (2008) based their study instead on actual admissions and administrative data from Texas public and private universities. Utilizing logistic regression models, they simulated annual admissions cohorts post-bans and *Hopwood* using counterfactual scenarios such as "affirmative action policies continued to be implemented" and "universities complied perfectly/automatically with *Hopwood*". Comparing these counterfactual models to actual admissions data revealed dramatic increases in admission rates for White students at all programs, increases for Asian students at selective programs, and steep decreases for Black and Hispanic students in all programs. We felt that Long & Tienda's approach was methodologically sound, and it influenced our decision to examine the admissions data using similar logistic regression and simulation approaches. Notably, Long & Tienda also discuss university-specific efforts to identify proxy admissions criteria (such as economic hardship, English as a second language, etc.) in order to "claw back" some of the lost diversity in admissions cohorts without specific reference to race or ethnicity. The data suggest that UT Austin may have successfully regained some ground in this effort, although nowhere near their original levels.

While not part of our study, studies such as **Hinrichs (2012)** have explored similar unanticipated effects in response to affirmative action bans, such as the "gaming" of AAP-impacted school districts by non-minority families to gain a competitive edge.

These findings illustrate the continued push and pull between legislatures, universities, and the broader public around access to educational and economic opportunities for all members of our society - and the debate is probably far from over.

Methodology

Our objective is to measure the effects of Affirmative Action bans and the Top-10% law on admissions to Texas A&M University (TAM). We downloaded the Texas Higher Education Opportunity dataset, available upon request for academic uses, which assigned a unique identifying number to students as they applied to various universities and as they progressed through their education. We limited our analysis to Texas A&M to avoid variances present in the data across multiple schools (i.e. race/ethnicity reporting, GPA scaling, etc), and TAM presented a notable drop in admission rates after the Top 10% law went into effect, thus providing a useful exemplar for building our analysis. Moreover, while some colleges only provided data prior to or after 1998, TAM records included years prior to and following the judicial ban on affirmative action. This was especially significant because, while the judicial restriction applied to all schools in the 5th Circuit District, the top 10% policy was restricted to public colleges and universities. We primarily used demographic and academic predictors, employing records from 1992-2022 that included information about admissions selectivity, public/private status, and the ethno-racial composition of their student body. These records contained a plethora of information about the applicant pool, and have been standardized where necessary, and checked for consistency (Table 1). The application dataset contained 163,027 observations of 24 predictor variables where each record represented an individual applicant (Table 2). These variables describe items typically found on a college admissions application such as the year and term an applicant desired to enroll, applicant demographics, applicant academic characteristics, and high school characteristics. Unfortunately, the data does not often include information regarding a student's high school academics or application essays, which is a notable constraint in attribution.

The target utilized for each record in the application dataset was the response variable `admit` (Institution's admission decision), a boolean where "1" indicated the person was admitted. Some of the predictor variables were transformed to binary to work with the missing values. As most of the variables were of a categorical nature, no other transformations were necessary. The resulting dataset is shown in Table 3. The probit generalized linear model was the base for the predictions since our dependent variable was binary (0 and 1). To improve the model's performance, stepwise selection and Lasso regression were applied to the original baseline model. The models were trained on the data that included applicants prior to Fall 1998 (before the Top-10% law). This model was then applied to predict admission status for all applicants, year by year, starting in the Fall of 1998 and compared against the actual admissions counts to identify changes in the admission process attributable to the law. In other words, in the case that our model, which was trained on the data *without* Hopwood's law, predicted admission with the law with high accuracy, we could state that we did not see any real changes in the admission process attributable to Hopwood. On the other hand, if the model's predictions differed from the data after the Fall of 1998, the law would be presumed to have had an effect.

Experimentation and Results

Data & Transformations

During our data exploration, we identified 21 categorical variables and 3 numeric variables (Table 1). Notably, the variables associated with ACT composite score, graduation year, and participation in the Century or Longhorn scholarship program had more than 50% missing data, and therefore, we removed them (Table 4). Our analysis revealed that 76% of all applicants in the dataset were admitted to Texas A&M University (Figure 1). Further investigation into the variables related to ethnicity, citizenship, Texas residency, and sex showed that over 50% of the data pertained to applicants who were US citizens, Texas residents, and identified as White/non-Hispanic (Figure 2).

To enhance the performance of our model, we transformed certain variables into binary form. Specifically, we converted sex to a binary indicator of whether a student is male or not, citizenship to indicate if a student is a US citizen, residency to indicate if a student is a Texas resident, and admit, and enroll, to represent whether a student was admitted (1/0) or enrolled (1/0). For columns containing information about each student's high school, decile, and quartile, missing values were replaced with a new level called "None." Additionally, students with missing values for the ethnicity variable were combined with those identified as "White, Non-Hispanic," in line with the Tienda study. We created a separate binary column to indicate whether a student was in the top 10% decile. Other variables with missing values were not considered in our research and were not imputed.

Thus, the final dataset for model training comprised the most essential applicant information, including admission/enrollment status, starting semester, sex, ethnicity, US citizenship/Texas residency, SAT score, year of admission, high school type (private or not), Texas state high school indicator, and top 10% decile (Table 5). The distribution of admission by ethnicity was found to be approximately equal across all ethnicities (Figure 3), and students with higher test scores were more frequently admitted (Figure 4).

Models

In our analysis, we initially constructed a generalized linear model using a dataset that included transformed categorical variables and spanned from 1992 to Fall 1998, consisting of 69,019 observations of 13 variables. As our dependent variable “admit” was binary (0 and 1), we employed logistic regression with a stepwise selection method, utilizing the glm() function with family=binomial. The variables outlined in Table 5 were included, and the model's outcomes are presented in Table 6 (AIC=50444.85, residual deviance=50412). The p-value associated with the Chi-Square Statistic was determined to be 0 (less than .05), indicating the potential usefulness of the model. Our findings indicated that ethnicity had the most negative impact on university acceptance, while Texas residency and inclusion in the top 10% decile positively influenced admission. Upon examining the residual plots (Figure 5), we observed under-fitting at lower predicted values, where the predicted proportions exceeded the observed proportions. The model performed reasonably well within the middle range but displayed extremes on the right and left sides. The normality of residuals in the binomial logistic regression models

served as evidence of a decent fit. The Standardized Residuals plot exhibited a constant variance, although some outliers were present. We assessed multicollinearity and found that all variables had a VIF (Variance Inflation Factor) of less than 5, indicating that multicollinearity was not a concern for our model. The confusion matrix revealed an accuracy of 87.7%. The model showed that over the years, factors such as Texas residency and being in the top 10% became more and more important (Figure 6, Table 7).

To enhance the model's performance, we opted to implement Lasso regression using the same dataset. The AIC value improved to -23258.91, and the importance of variables for the admission decision remained consistent in the new model: ethnicity, top 10%, and Texas residency were identified as the most significant features (Table 8). We extracted the coefficients using lambda.min, which minimizes the mean cross-validated error (AIC=-23258.91). The confusion matrix demonstrated an accuracy of 87.8%. The model showed that over the years being in the top 10% became more and more important (Figure 7, Table 9). The Standardized Residuals plot exhibited a constant variance, albeit with some outliers. By employing Lasso regression, we addressed the issue of multicollinearity by selecting the variable with the largest coefficient while setting the remaining variables to (almost) zero (Figure 8). Consequently, due to its superior performance in terms of the AIC value, we utilized the Lasso model for further analysis.

Predictions & Results

In this study, we utilized the Lasso regression model trained on pre-Fall 1998 data to investigate the impact of various factors on admission outcomes at Texas A&M University. Our goal was to assess whether the university had indeed altered the importance assigned to applicant characteristics, specifically with regards to underrepresented minority applicants, or if ethnicity continued to be a significant factor in admission decisions. To achieve this, we compared the admission outcomes of our simulated class to the actual data, enabling us to infer the net effect of factors such as the Hopwood decision and the top-10% policy.

However, upon examining the actual admission demographics of Texas A&M University (Table 10, Figure 9), we found that they did not align with our predicted values. The percentage of admitted students from the White ethnic group actually increased, while the percentage of Black, Hispanic, and Asian ethnic groups decreased, relative to our predictions.

Figure 10 and Table 11 show the disparities between our model predictions (from 1998 to 2022 in the absence of a ban) and the actual data. We observed that the model underestimated the number of white/non-Hispanic students accepted after Hopwood, while overestimating the number of Hispanic and Black students. Interestingly, before Hopwood, there appeared to be a preference for underrepresented minority students; however, after Hopwood, Black and Hispanic applicants were less likely to be accepted compared to their White counterparts. This finding suggests that, contrary to expectations, the top 10% law did not result in increased acceptance rates for students from minority ethnicities. Indeed, the model also indicated that the most influential factor for admission was whether a student belonged to the top 10%, or was a Texas resident, and de-emphasized the importance of ethnicity altogether.

Discussion and Conclusions

Overall, our analysis aimed to highlight how admissions outcomes have changed at a major Texas public university after the implementation of *Hopwood* and the Top 10% mandate. Our analysis provides a quantifiable basis for highlighting these changes, and suggests that these changes led to a decrease in diversity of ethnicities in admissions, at least at Texas A&M. Instead, *Hopwood* and the Top 10% law led to ever increasing importance of students' status in the top decile of their graduating high school class, as well as being Texas residents. The idea that the Top 10% law would reduce competition among applicants may have been true, but only among those who were already top performers in their high schools. It has been well studied that such performers tend to belong to less diverse ethnicities to begin with, and so we can surmise that the law's effects were essentially to kick the onus of diversity initiatives upstream, to secondary schools. In this way, the top 10% law did not, in fact, provide a useful alternative to affirmative action. Although our numbers and discrepancies were relatively small, we were confident in our model, so we believe they accurately reflect the situation. Ultimately, the complex nature of the admission process, along with a lack of data into high school transcripts, financial aid data, or other data, means further study is warranted, but we hope this can serve as a useful basis for such study at other universities both within Texas and throughout the United States.

References

1) Card, D., & Krueger, A. B. (2005). Would the Elimination of Affirmative Action Affect Highly Qualified Minority Applicants? Evidence from California and Texas. *Industrial and Labor Relations Review*, *58*(3), 416–434. https://doi.org/10.1177/001979390505800306
2) Fiel, J. E. (2022). Opportunity Seeking Across Segregated Schools: Unintended Effects of Automatic Admission Policies on High School Segregation. *Educational Evaluation and Policy Analysis*, *44*(3), 485–504. https://doi.org/10.3102/01623737221078286
3) Hinrichs, P. (2012). The Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities. The Review of Economics and Statistics, 94(3), 712–722. https://doi.org/10.1162/rest_a_00170
4) Long, M. C., & Tienda, M. (2010). Changes in Texas universities' applicant pools after the Hopwood decision. *Social Science Research*, *39*(1), 48–66. https://doi.org/10.1016/j.ssresearch.2009.06.004
5) Long, M. C. (2004). Race and College Admissions: An Alternative to Affirmative Action? *The Review of Economics and Statistics*, *86*(4), 1020–1033. https://doi.org/10.1162/0034653043125211
6) Long, M. C., & Tienda, M. (2008). Winners and Losers: Changes in Texas University Admissions Post-*Hopwood*. *Educational Evaluation and Policy Analysis*, *30*(3), 255–280. https://doi.org/10.3102/0162373708321384

Appendix

Tables

Table 1: Variable Availability

Variable Name	Variable Label	Category
studentid	Student ID	Metadata
yeardes	Year admission desired	Application Identifiers
termdes	Term admission desired	Application Identifiers
male	Male	Demographic Characteristics
ethnic	Ethnicity	Demographic Characteristics
citizenship	Citizenship	Demographic Characteristics
restype	Resident type	Demographic Characteristics
satR*	SAT composite score (Recentered)	Academics
actR	ACT composite score	Academics
testscoreR*	SAT if provided, or ACT set to SAT Scale Student HS class rank: decile	Academics
decileR	Student HS class rank: decile	Academics
quartile	Student HS class rank: quartile	Academics
major_field	Field of desired first choice major	Academics
hsprivate	Private HS	High School
hstypeR	HS type: regular or other	High School
hsinstate	Texas HS	High School
hseconstatus	HS economic status	High School
hseconstatus	HS economic status	High School
hslos	HS in Longhorn Scholarship program	High School
hscentury	HS in Century Scholarship program	High School
admit	Admitted	Admissions
admit_prov	Provisional admission	Admissions
enroll	Enrolled	Admissions
gradyear	Graduation year	Admissions

Table 2: Application Dataset

studentid	yeardes	termdes Term admission desired	male	ethnic	satR	testscoreR	decileR Student HS class rank: decile - RECODE	quartile Student HS class rank: quartile	major_field Field of desired 1st choice major	hsprivate Private HS
000004	1998	Fall	0	White, Non-Hispanic	38	38	Second Decile	Top 25%	AGRICULTURE	Public
000005	1998	Fall	1	White, Non-Hispanic	20	20	Fifth Decile	Second quartile	NATURAL/PHYSICAL SCIENCES	Public
000006	1998	Fall	1	White, Non-Hispanic	66	66	Top 10%	Top 25%	ENGINEERING/COMPUTER SCIENCE	Public
000007	1998	Fall	1	White, Non-Hispanic	27	27	Third Decile	Second quartile	AGRICULTURE	Public
000008	1998	Fall	1	White, Non-Hispanic	86	86	Second Decile	Top 25%	NATURAL/PHYSICAL SCIENCES	Public
000009	1998	Fall	0	White, Non-Hispanic	58	58	Fifth Decile	Second quartile	NATURAL/PHYSICAL SCIENCES	Public
000011	1998	Fall	1	Other	56	56	Fourth Decile	Second quartile	SOCIAL SCIENCES	Public

Table 3: Transformed dataset

admit	termdes Term admission desired	male	ethnic	US_Citizen	Texas_resident	satR	testscoreR	top10	hsprivate Private HS	hsinstate Texas HS	yeardes	enroll
1	Fall	0	White, Non-Hispanic	1	1	38	38	FALSE	Public	Yes	1998	0
0	Fall	1	White, Non-Hispanic	1	1	20	20	FALSE	Public	Yes	1998	0
1	Fall	1	White, Non-Hispanic	1	1	66	66	TRUE	Public	Yes	1998	0
1	Fall	1	White, Non-Hispanic	1	1	27	27	FALSE	Public	Yes	1998	1
1	Fall	1	White, Non-Hispanic	1	1	86	86	FALSE	Public	Yes	1998	0
1	Fall	0	White, Non-Hispanic	1	1	58	58	FALSE	Public	Yes	1998	0
1	Fall	1	Other	1	1	56	56	FALSE	Public	Yes	1998	1
1	Fall	0	Asian or Pacific Isla...	1	1	38	38	FALSE	Private	Yes	1998	0

Table 4: Application Data: missing values

variable	complete_rate	n_missing	min	max
actR	0.46	72133	1	24
admit	1.00	0	NA	NA
admit_prov	1.00	0	NA	NA
decileR	1.00	0	NA	NA
enroll	1.00	0	NA	NA
ethnic	1.00	50	NA	NA
gradyear	0.35	86914	2	22
hscentury	0.32	91636	NA	NA
hseconstatus	0.72	38094	NA	NA
hsinstate	1.00	103	NA	NA
hslos	0.41	79635	NA	NA
hsprivate	0.98	2375	NA	NA
hstypeR	0.91	12338	NA	NA
major_field	1.00	2	NA	NA
male	1.00	52	NA	NA
quartile	1.00	10	NA	NA
satR	1.00	0	3	111
studentid	1.00	0	NA	NA
studentid_uniq	1.00	0	NA	NA
termdes	1.00	0	NA	NA
testscoreR	1.00	0	3	111
Texas_resident	1.00	0	NA	NA
US_Citizen	1.00	0	NA	NA
yeardes	1.00	0	1992	2002

Table 5: Dataset for the model

variable	complete_rate	n_missing	min	max
admit	1	0	NA	NA
enroll	1	0	NA	NA
ethnic	1	0	NA	NA
hsinstate	1	0	NA	NA
hsprivate	1	0	NA	NA
male	1	0	NA	NA
satR	1	0	3	111
termdes	1	0	NA	NA
testscoreR	1	0	3	111
Texas_resident	1	0	NA	NA
top10	1	0	NA	NA
US_Citizen	1	0	NA	NA
yeardes	1	0	1992	2002

Table 6: Baseline model

Observations	69019
Dependent variable	admit
Type	Generalized linear model
Family	binomial
Link	probit

χ(15)	23390.53
Pseudo-R (Cragg-Uhler)	0.44
Pseudo-R (McFadden)	0.32
AIC	50444.85
BIC	50591.12

	Est.	S.E.	z val.	p
(Intercept)	-1.06	0.06	-18.70	0.00
male1	-0.14	0.01	-10.98	0.00
ethnicHispanic	-0.03	0.04	-0.70	0.49
ethnicAmerican Indian/Alaskan Native	-0.88	0.09	-9.36	0.00
ethnicAsian or Pacific Islander	-0.92	0.04	-21.49	0.00
ethnicInternational	-0.48	0.09	-5.41	0.00
ethnicOther	-0.99	0.07	-14.35	0.00
ethnicWhite, Non-Hispanic	-0.80	0.03	-24.51	0.00
US_Citizen1	0.08	0.04	1.82	0.07
Texas_resident1	0.73	0.03	21.83	0.00
satR	0.03	0.00	79.97	0.00
top10TRUE	1.47	0.02	75.05	0.00
hsprivatePrivate	-0.17	0.03	-6.65	0.00
hsprivateNone	-0.28	0.05	-6.02	0.00
hsinstateYes	0.07	0.04	1.81	0.07
hsinstateNone	-0.21	0.20	-1.01	0.31

Table 7: GLM model's coefficients, 1992-2002

glm() Admission Coefficients

coef	< 1998	1998	1999	2000	2001	2002
(Intercept)	-1.065	-1.437	-1.719	-1.277	-1.408	-1.699
male1	-0.144	-0.399	-0.252	-0.145	-0.114	-0.087
ethnicHispanic	-0.026	0.117	-0.116	-0.100	0.074	0.173
ethnicAmerican Indian/Alaskan Native	-0.879	0.346	0.095	-0.047	0.081	0.089
ethnicAsian or Pacific Islander	-0.918	-0.194	-0.383	-0.407	-0.421	-0.211
ethnicInternational	-0.485	-0.181	0.069	-0.061	-0.293	-0.114
ethnicOther	-0.989	-0.300	0.023	-0.367	-0.369	-0.184
ethnicWhite, Non-Hispanic	-0.798	0.204	0.070	-0.246	0.058	0.181
US_Citizen1	0.077	NA	0.251	NA	NA	NA
Texas_resident1	0.733	0.836	0.514	0.627	0.376	0.574
satR	0.028	0.031	0.019	0.021	0.021	0.021
top10TRUE	1.473	2.001	1.916	2.591	3.052	3.276
hsprivatePrivate	-0.171	-0.162	-0.157	NA	-0.123	-0.242
hsprivateNone	-0.285	-0.487	-0.144	NA	-0.087	-0.185
hsinstateYes	0.065	NA	0.443	-0.210	NA	NA
hsinstateNone	-0.206	NA	0.388	0.326	NA	NA

Table 8: Lasso regression model

	Est
top10TRUE	2.740
Texas_resident1	1.244
US_Citizen1	0.120
hsinstateYes	0.084
satR	0.048
ethnicHispanic	0.000
testscoreR	0.000
male1	-0.246
hsprivatePrivate	-0.281
hsinstateNone	-0.282
hsprivateNone	-0.485
ethnicInternational	-0.796
ethnicWhite, Non-Hispanic	-1.322
ethnicAmerican Indian/Alaskan Native	-1.423
ethnicAsian or Pacific Islander	-1.537
ethnicOther	-1.599
(Intercept)	-1.834

Table 9: Lasso model's coefficients, 1992-2002

Lasso Admission Coefficients

coef	< 1998	1998	1999	2000	2001	2002
top10TRUE	2.740	3.899	3.399	4.899	5.815	6.375
Texas_resident1	1.244	1.225	0.791	0.820	0.510	0.968
US_Citizen1	0.120	-0.242	0.436	0.051	0.174	-0.044
hsinstateYes	0.084	0.312	0.581	-0.245	0.117	-0.064
satR	0.048	0.056	0.027	0.029	0.028	0.027
male1	-0.246	-0.706	-0.365	-0.199	-0.130	-0.108
hsprivatePrivate	-0.281	-0.229	-0.201	-0.066	-0.188	-0.348
hsinstateNone	-0.282	NA	NA	0.431	-0.340	0.527
hsprivateNone	-0.485	-0.700	-0.179	-0.115	-0.093	-0.326
ethnicInternational	-0.796	-0.701	NA	-0.043	-0.360	-0.088
ethnicWhite, Non-Hispanic	-1.322	0.341	0.100	-0.349	0.012	0.266
ethnicAmerican Indian/Alaskan Native	-1.423	0.579	NA	NA	0.003	0.061
ethnicAsian or Pacific Islander	-1.537	-0.376	-0.590	-0.567	-0.690	-0.353
ethnicOther	-1.599	-0.510	NA	-0.483	-0.442	-0.304
(Intercept)	-1.834	-2.420	-2.441	-1.741	-2.071	-2.299
ethnicHispanic	NA	0.149	-0.101	NA	0.111	0.332

Table 10: Actual Admission Demographics for Texas A&M

Admission Demographics Actual

ethnic	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Black, Non-Hispanic	0.043	0.046	0.053	0.057	0.049	0.037	0.035	0.036	0.033	0.035	0.034
Hispanic	0.127	0.139	0.153	0.155	0.133	0.114	0.105	0.101	0.121	0.120	0.112
American Indian/Alaskan Native	0.004	0.003	0.004	0.004	0.004	0.005	0.005	0.005	0.005	0.005	0.004
Asian or Pacific Islander	0.064	0.053	0.058	0.052	0.048	0.061	0.056	0.058	0.063	0.060	0.066
International	0.005	0.005	0.004	0.004	0.003	0.003	0.004	0.004	0.006	0.005	0.004
Other	0.002	0.002	0.002	0.010	0.012	0.023	0.014	0.014	0.021	0.009	0.004
White, Non-Hispanic	0.755	0.753	0.725	0.718	0.751	0.756	0.782	0.782	0.751	0.765	0.776

Table 11: The difference between the predicted admission and actual

Admission Demographics (predicted - actual)

ethnic	1998	1999	2000	2001	2002
Black, Non-Hispanic	0.003	0.005	0.002	0.006	0.008
Hispanic	0.006	0.015	0.006	0.008	0.011
American Indian/Alaskan Native	0.000	0.000	0.000	0.000	0.000
Asian or Pacific Islander	0.001	0.004	0.000	0.005	0.004
International	0.000	0.000	-0.001	0.000	0.000
Other	0.000	-0.001	-0.002	0.000	0.000
White, Non-Hispanic	-0.009	-0.022	-0.004	-0.020	-0.023

Figures

Figure 1: Admission rate 1992-2002

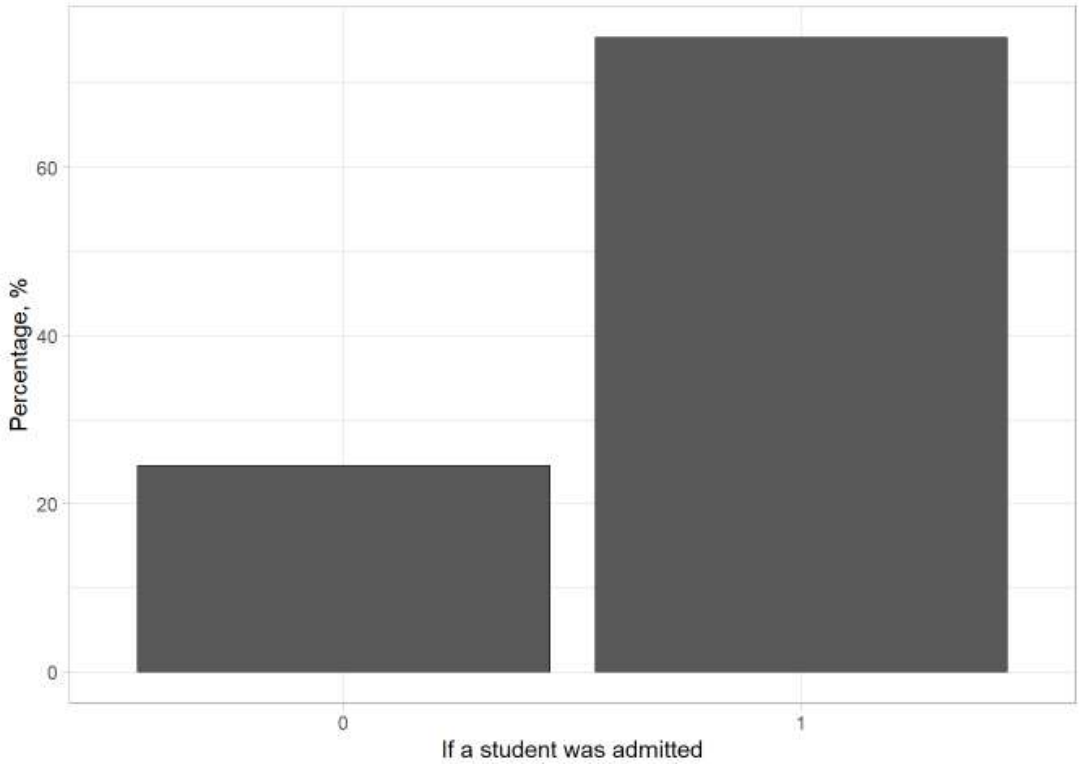


Figure 2: Applications' diversity

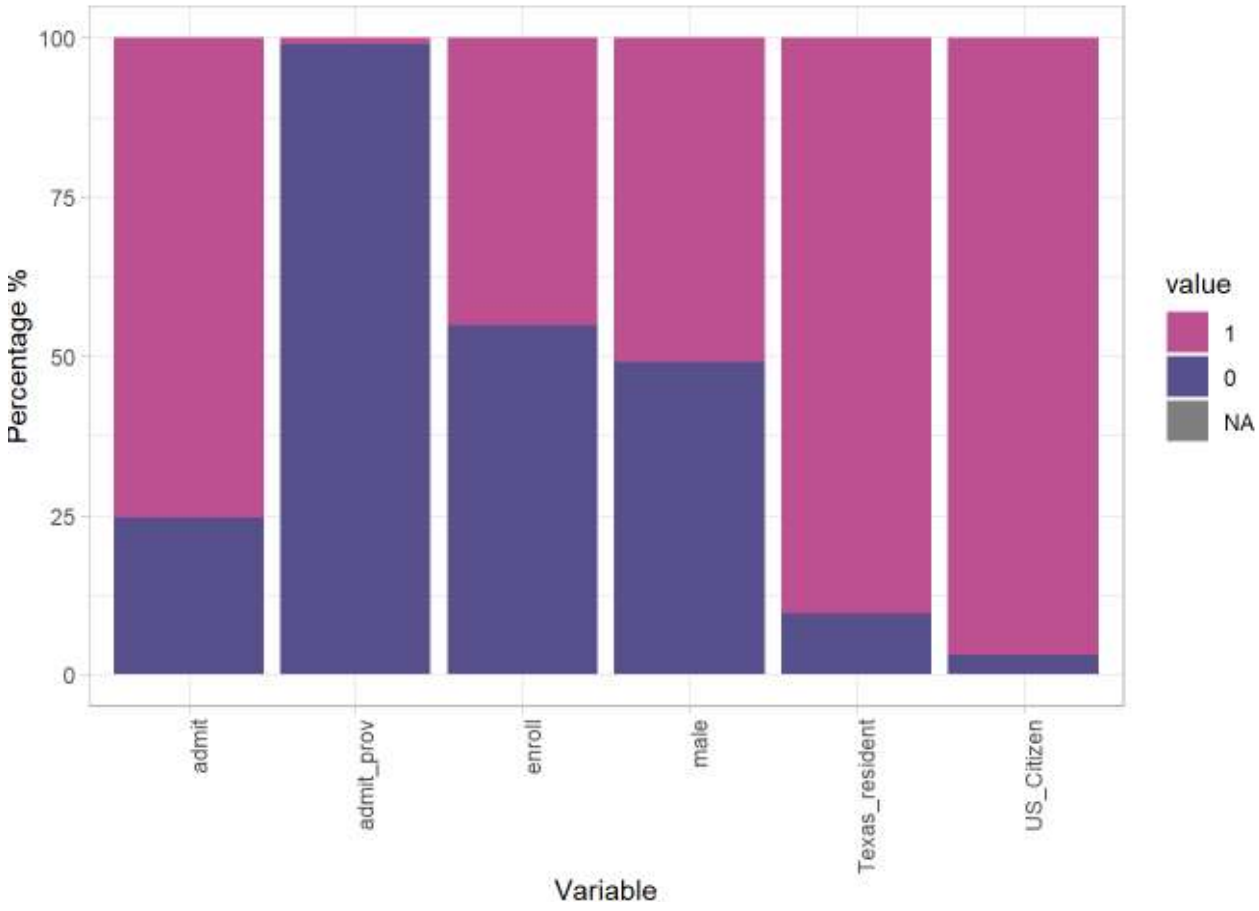


Figure 3: Admission rate by ethnicity

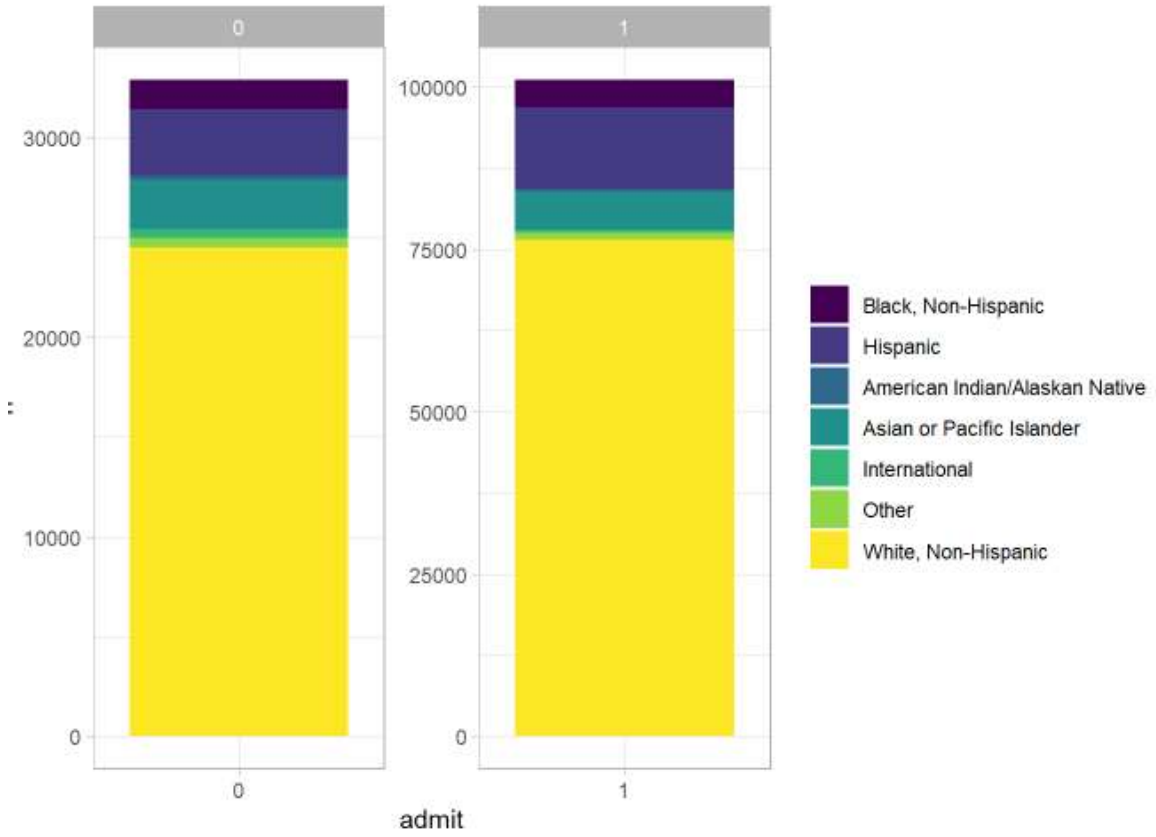


Figure 4: Admission/Enrollment vs. Test scores

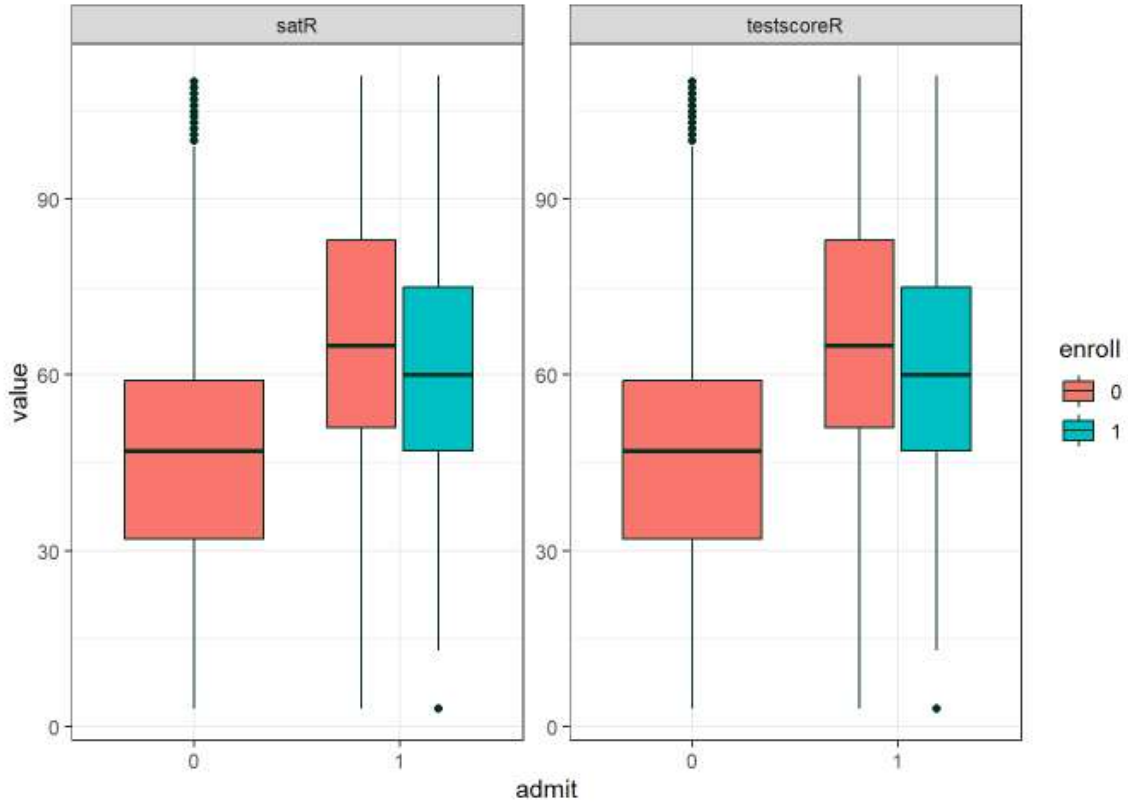


Figure 5: Baseline model's assumptions

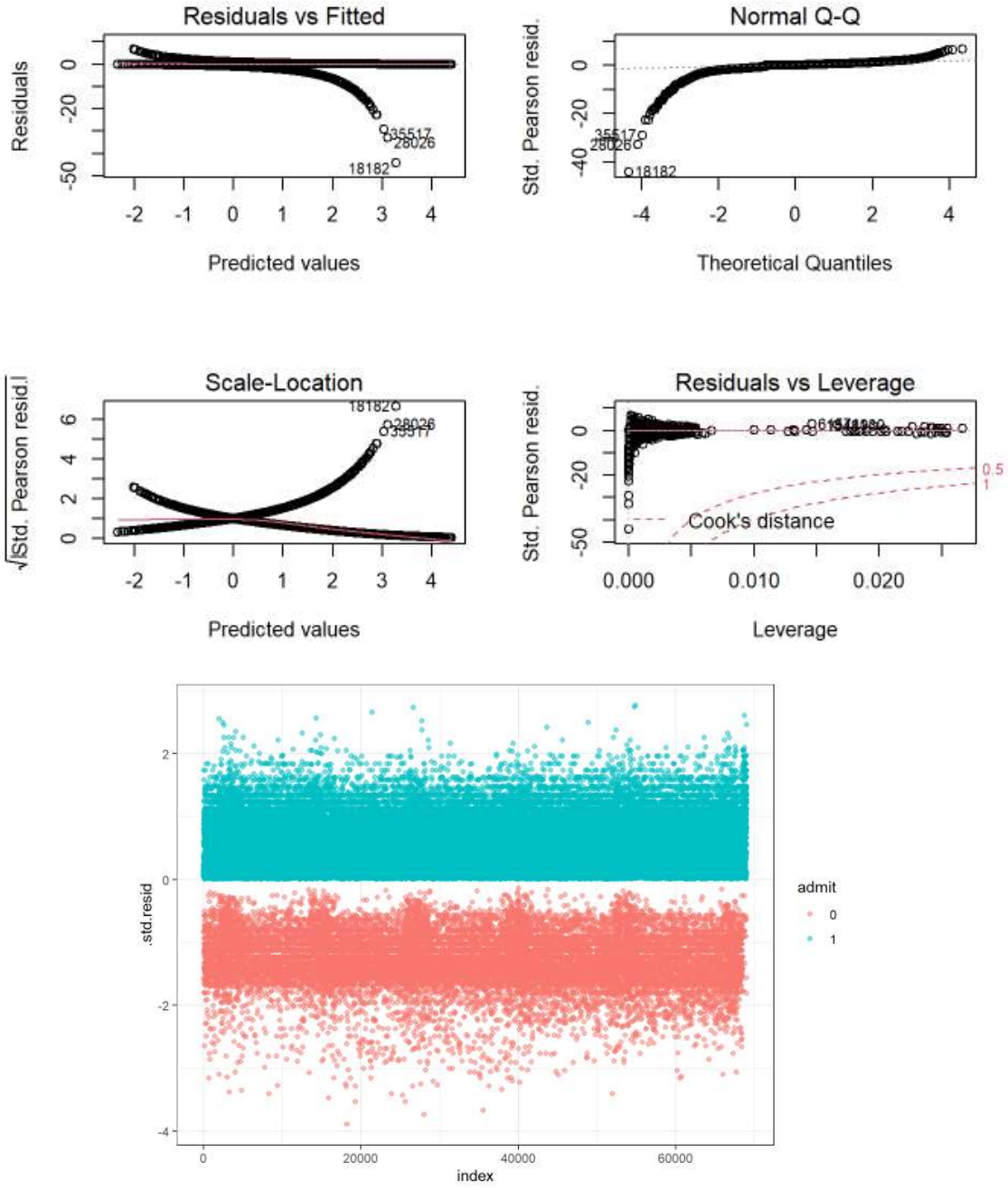


Figure 6: GLM model's coefficients, 1992-2002

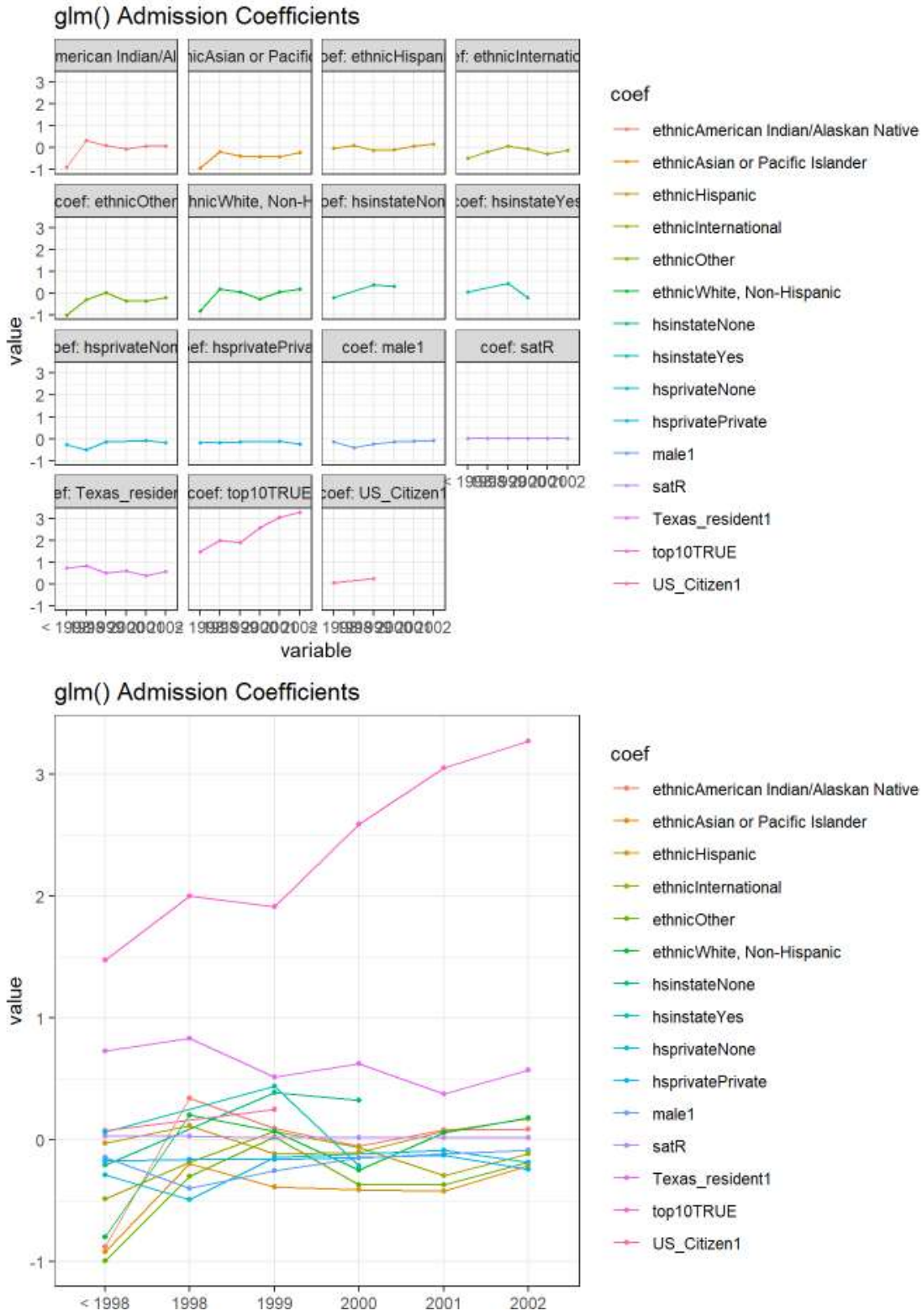
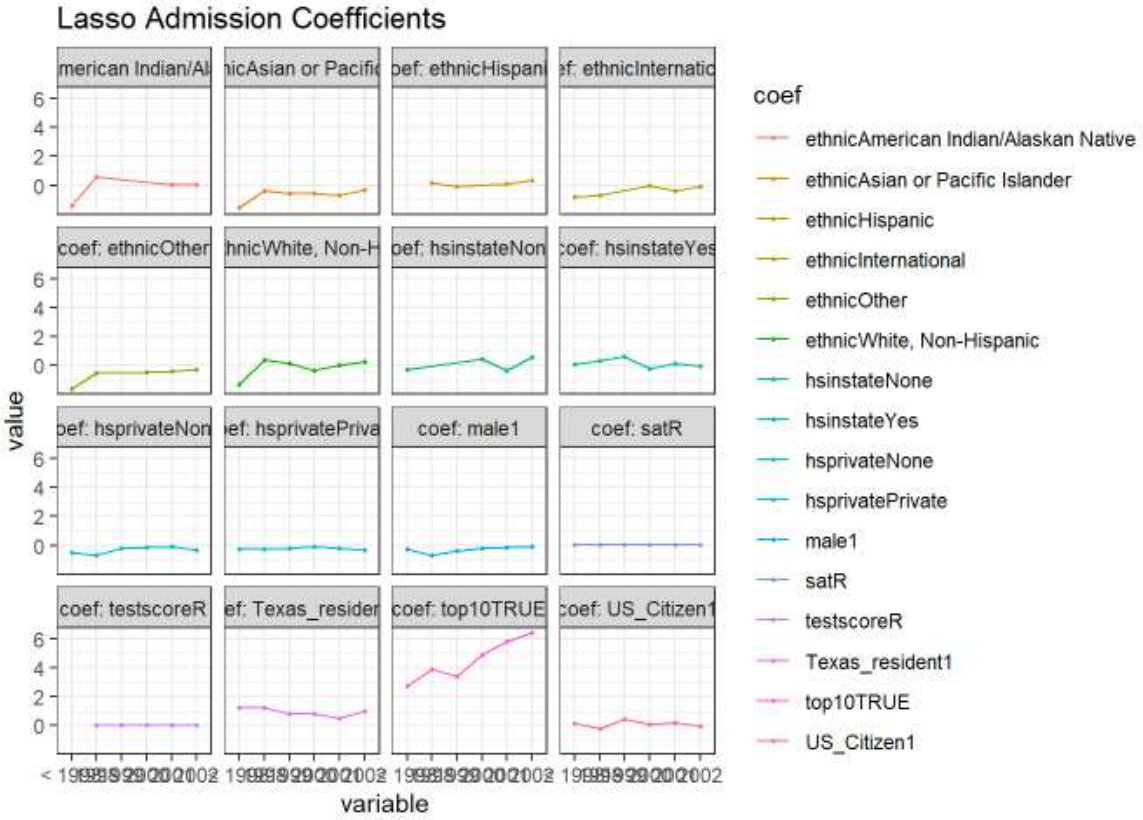


Figure 7: Lasso model's coefficients, 1992-2002



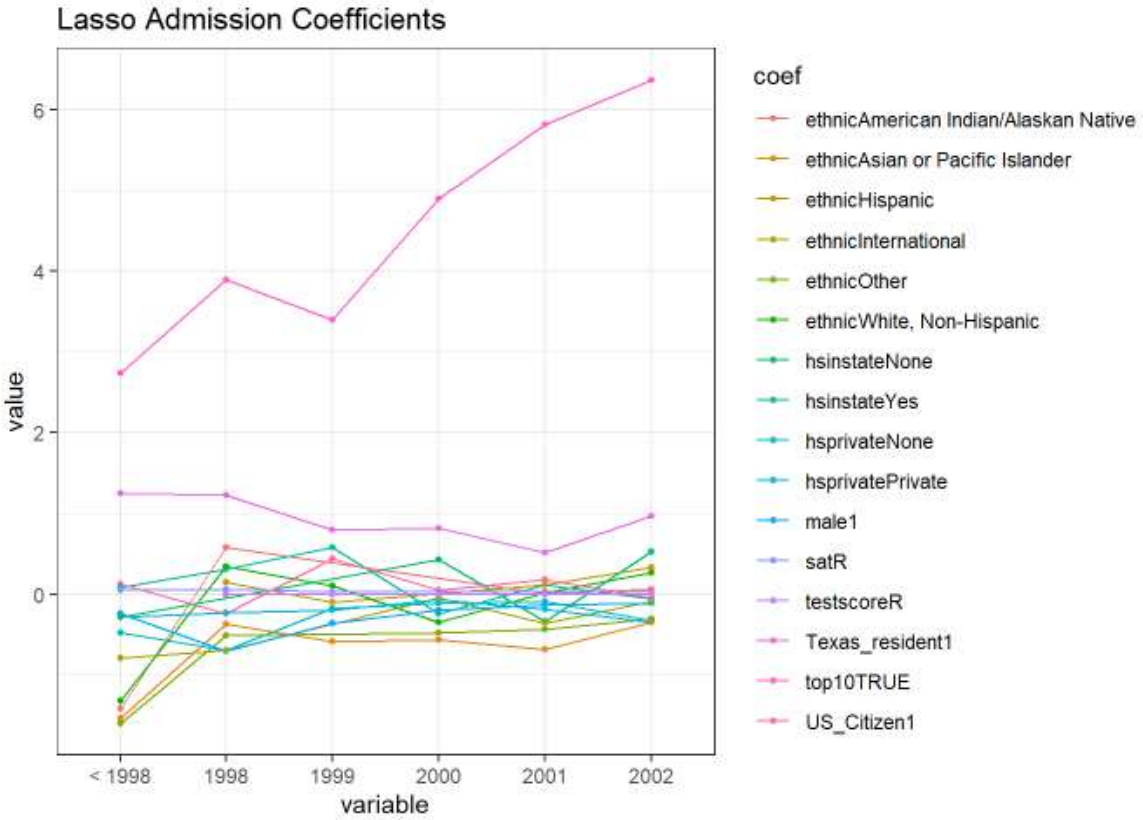


Figure 8: Lasso model's assumptions

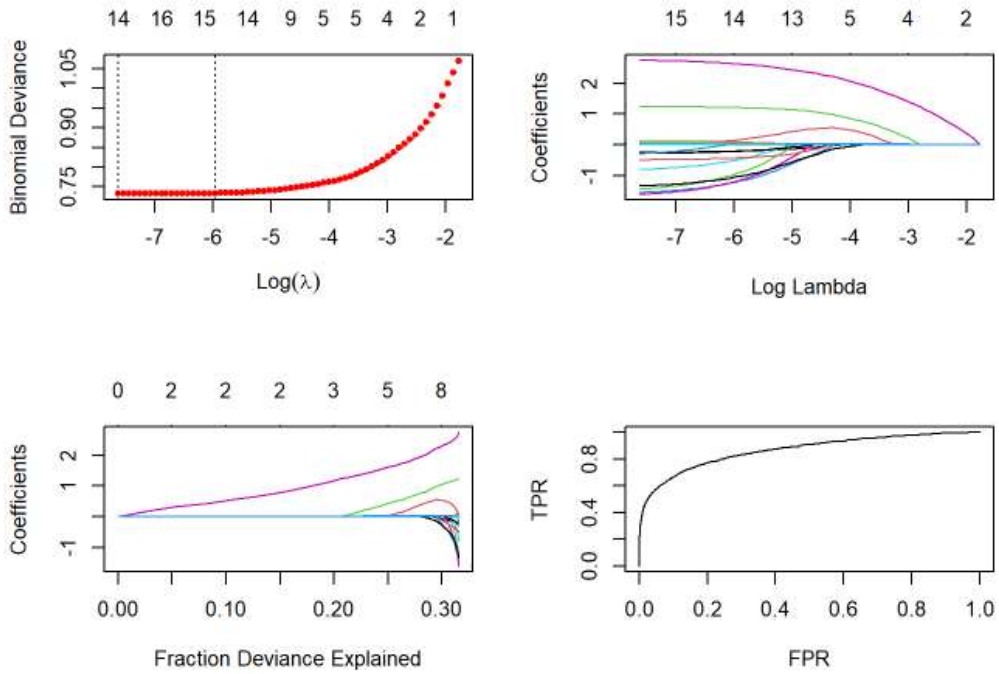


Figure 9: Actual Admission Demographics for Texas A&M

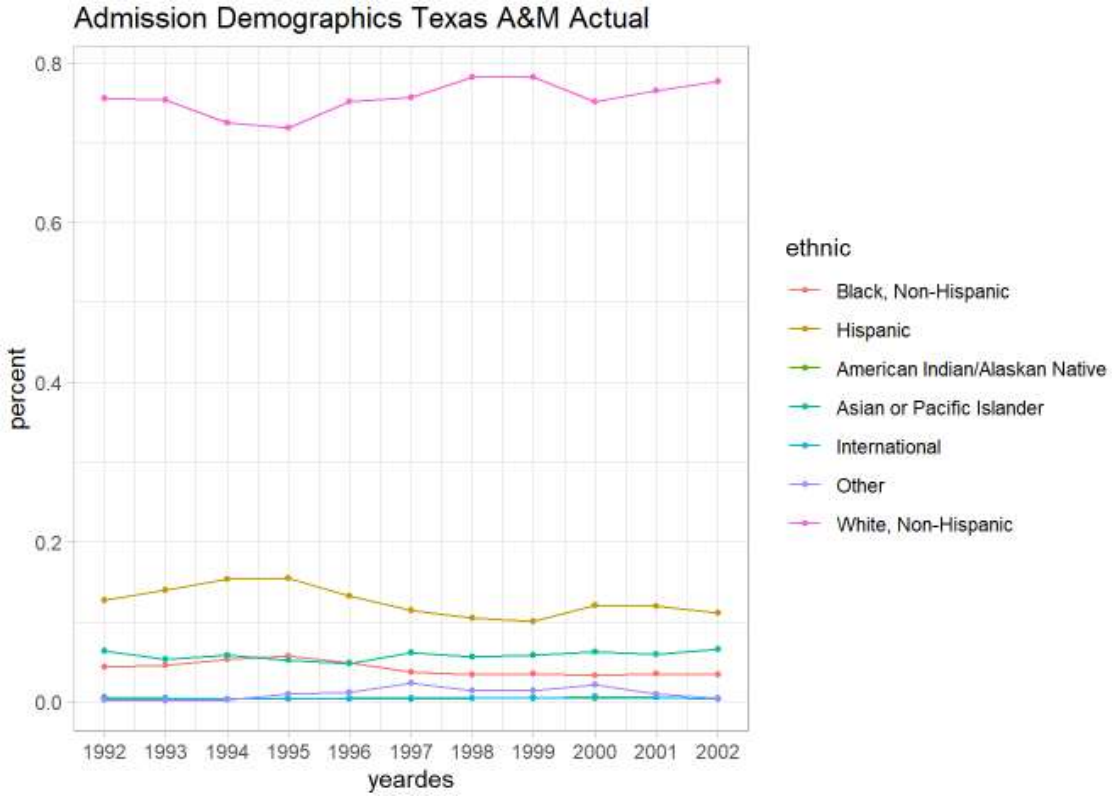
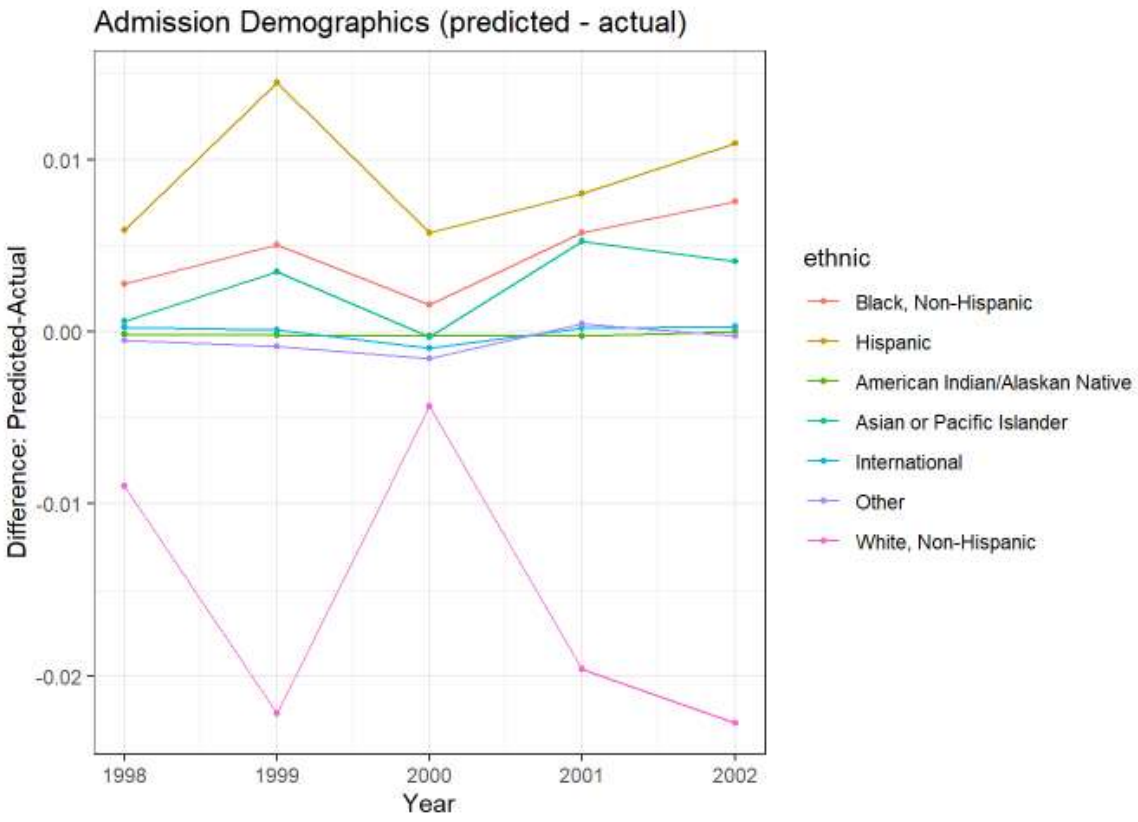


Figure 10: The difference between the predicted admission and actual



R Code

(Setup)

Chunk 1

```
{r setup, include=FALSE}
# chunks
knitr::opts_chunk$set(
            eval = TRUE,
            echo = FALSE,
            fig.align = "center",
            fig.height = 5,
            message = FALSE,
            warning = FALSE,
            include = TRUE
)
# libraries
library(tidyverse)
library(kableExtra)
library(MASS) # glm.nb()
library(mice)
library(pscl) # zeroinfl()
library(skimr)
library(sjPlot)
library(mpath)
library(yardstick)
library(labelled)
library(haven)
library(corrplot)
library(Hmisc)
library(jtools)
library(caret)
library(broom)

 # library
library(treemap)
library(ggplot2)
library(hrbrthemes)
library(viridis)
library(vip)
```

```
library(ggpubr)
library(rockchalk)

# ggplot
theme_set(theme_light())

# random seed
set.seed(42)
```

Chunk 2

```
{r common functions}

#' nice_table
#'
#' @param df
#' @param fw
nice_table <- function(df, cap=NULL, cols=NULL, dig=3, fw=F){
  if (is.null(cols)) {c <- colnames(df)} else {c <- cols}
  table <- df %>%
    kable(caption=cap, col.names=c, digits=dig) %>%
    kable_styling(
      bootstrap_options = c("striped", "hover", "condensed"),
      html_font = 'monospace',
      full_width = fw)
  return(table)
}

model_diag <- function(model){
  model_sum <- summary(model)
  aic <- AIC(model)
  ar2 <- model_sum$adj.r.squared
  disp <- sum(resid(model,'pearson')^2)/model$df.residual
  loglik <- logLik(model)

  vec <- c(ifelse(is.null(aic), NA, aic),
           ifelse(is.null(ar2), NA, ar2),
           ifelse(is.null(disp), NA, disp),
           ifelse(is.null(loglik), NA, loglik))

  names(vec) <- c('AIC','Adj R2','Dispersion','Log-Lik')
  return(vec)
}
```

```
factor_haven <- function(df, col_lst) {
  for (c in col_lst) {
    df[c] <- as_factor(zap_missing(df[c]))
  }
  return(df)
}


#' Title
#'
#' @param fit
#' @param lambda
#'
#' @return
#' @export
#'
#' @examples
glmnet_cv_aicc <- function(fit, lambda = 'lambda.1se'){
  whlm <- which(fit$lambda == fit[[lambda]])
  with(fit$glmnet.fit,
       {
         tLL <- nulldev - nulldev * (1 - dev.ratio)[whlm]
         k <- df[whlm]
         n <- nobs
         return(list('AICc' = - tLL + 2 * k + 2 * k * (k + 1) / (n - k - 1),
                     'BIC' = log(n) * k - tLL))
       })
}


#' coeff2dt
#'
#' @param fitobject
#' @param s
#'
#' @return
#' @export
#'
#' @examples
coeff2dt <- function(fitobject, s) {
  coeffs <- coef(fitobject, s)
  coeffs.dt <- data.frame(name = coeffs@Dimnames[[1]][coeffs@i + 1], coefficient =
coeffs@x)
```

```
  # reorder the variables in term of coefficients
  return(coeffs.dt[order(coeffs.dt$coefficient, decreasing = T),])
}


#' Title
#'
#' @param df
#' @param y
#' @param c_df
#'
#' @return
#' @export
#'
#' @examples
glm_coef <- function(df, y, c_df) {

  am_df <- df %>% dplyr::filter(yeardes == y & termdes=="Fall") %>%
dplyr::select(!c(enroll,yeardes,termdes))

  am_model <- glm(am_df, formula = admit ~ . , family = binomial(link = "probit"))
  summary(am_model)

  am_model_aic <- am_model %>% stepAIC(trace = FALSE)
  summ(am_model_aic)

  a <- as.data.frame(summary(am_model_aic)$coefficients[,1])
  a <- cbind(coef = rownames(a), a)
  rownames(a) <- NULL
  names(a) <- c('coef',y)
  a$model <- 'glm'

  if (is.null(c_df)) {
    c_df <- a
  } else {
    c_df <- c_df %>% full_join( a, by=c('coef','model'))
  }

  return (c_df)
}


#' Title
#'
#' @param df
```

```
#' @param y
#' @param c_df
#'
#' @return
#' @export
#'
#' @examples
lasso_coef <- function(df, y, c_df) {

  if (y == 1998) {
    am_df <- df %>% dplyr::filter(yeardes == y & termdes=="Fall") %>%
dplyr::select(!c(enroll,yeardes,termdes))
  } else {
    am_df <- df %>% dplyr::filter(yeardes == y) %>%
dplyr::select(!c(enroll,yeardes,termdes))
  }

  X <- model.matrix(admit ~ . , data=am_df)[,-1]
  Y <- am_df["admit"]


  lasso.model<- cv.glmnet(x=X,y=Y,
                          family = "binomial",
                          link = "probit",
                          standardize = TRUE,                    #standardize
                          nfold = 5,
                          alpha=1)                               #alpha=1 is lasso

  l.min <- lasso.model$lambda.min
  coef(lasso.model, s = "lambda.min" )


  a <- coeff2dt(fitobject = lasso.model, s = "lambda.min")
  names(a) <- c('coef',y)
  a$model <- 'lasso'

  if (is.null(c_df)) {
    c_df <- data.frame(a)
  } else {
    c_df <- c_df %>% full_join( a, by=c('coef','model'))
  }


  return (c_df)
}
```

```
#' Title
#'
#' @param df
#' @param model
#' @param y
#' @param d_df
#'
#' @return
#' @export
#'
#' @examples
demographicCount <- function(df, model, y, d_df) {

  if (y == 1998) {
    am_df <- df %>% dplyr::filter(yeardes == y & termdes=="Fall") %>%
dplyr::select(!c(enroll,yeardes,termdes))
  } else {
    am_df <- df %>% dplyr::filter(yeardes == y) %>%
dplyr::select(!c(enroll,yeardes,termdes))
  }

  am_pred <- predict.glm(model, am_df, "response")
  am_df$admit_prob <- am_pred
  am_df$admit_pred <- ifelse(am_pred >= 0.5, 1, 0)

  log_matrix_1 <- confusionMatrix(factor(am_df$admit_pred),
                                  factor(am_df$admit), "1")

  log_matrix_1


  d <- am_df %>%
    group_by(admit_pred,ethnic) %>%
    summarise(y = n())

  names(d) <- c('admit', 'ethnic', y)
  d_df <- d_df %>% full_join( d, by=c('admit','ethnic'))

  return(d_df)
}


#' Title
#'
```

```
#' @param df
#' @param l_model
#' @param y
#' @param d_df
#'
#' @return
#' @export
#'
#' @examples
demographicCountLasso <- function(df, l_model, y, d_df) {


  if (y == 1998) {
    am_df <- df %>% dplyr::filter(yeardes == y & termdes=="Fall") %>%
dplyr::select(!c(enroll,yeardes,termdes))
  } else {
    am_df <- df %>% dplyr::filter(yeardes == y) %>%
dplyr::select(!c(enroll,yeardes,termdes))
  }


  X_test <- model.matrix(admit ~ . ,data=am_df)[,-1]
  Y_test <- am_df[,"admit"]


  # predict using coefficients at lambda.min
  lassoPred <- predict(l_model, newx = X_test, type = "response", s = 'lambda.min')

  #pred_df <- am_lasso1998_df
  am_df$admit_prob <- lassoPred[,1]
  am_df$admit_pred <- ifelse(lassoPred >= 0.5, 1, 0)[,1]


  log_matrix_1 <- confusionMatrix(factor(am_df$admit_pred),
                                  factor(am_df$admit), "1")
  log_matrix_1


  d <- am_df %>%
    group_by(admit_pred,ethnic) %>%
    summarise(y = n())

  names(d) <- c('admit', 'ethnic', y)
  d_df <- d_df %>% full_join( d, by=c('admit','ethnic'))
```

```
  return(d_df)

}
```

(Data)

Chunk 3

```
{r}

# change to your local data dir outside the repo
local_data_dir <- 'C:/Users/daria/Documents/theop'
#local_data_dir <- '../../data/theop'

# # load application and transactions data frames
load(paste0(local_data_dir, '/data_model/df_applications.RData'))
load(paste0(local_data_dir, '/data_model/df_transcripts.RData'))
```

Chunk 4

```
{r}
# clean application data and remove labels
app_df  <- df_applications

col_lst <-
c("termdes","male","ethnic","citizenship","restype","satR","actR","testscoreR","decileR","
quartile","major_field","hsprivate","hstypeR","hsinstate","hseconstatus","hslos","hscentur
y","admit","admit_prov","enroll","gradyear","studentid_uniq","univ","termapp","sat_not_re
centeredR","admit_ut_summer")

app_df <- factor_haven(app_df,col_lst)

# clean transcripts and remove labels
transcript_df <- df_transcripts

col_lst <- c("term","hrearn","term_major_dept","term_major_field")

transcript_df <- factor_haven(transcript_df,col_lst)
```

Chunk 5

```
## Choose Texas A&M college and compare the admissions policy at a university before
and after top 10%. Make variables categorical if needed.
```

```
texas_applications <- filter(app_df, univ == "am") %>% dplyr::select(!c(termapp,
sat_not_recenteredR, admit_ut_summer, univ)) %>%
        drop_na(satR,decileR)
#texas_transcripts <- filter(transcript_df, univ == "am") %>% dplyr::select(!univ)
```

Chunk 6

```
{r}
## Dummy Variables for factors with two levels
dummy_vars <- function(df){
  df %>%
    mutate(
      male = factor(ifelse(male == "Male", 1, 0)),
      US_Citizen = factor(ifelse(citizenship == "US Citizen",  1, 0)),
      Texas_resident = factor(ifelse(restype == "Texas Resident",  1, 0)),
      admit = factor(ifelse(admit == "Yes",  1, 0)),
      admit_prov = factor(ifelse(admit_prov == "Yes",  1, 0)),
      enroll = factor(ifelse(enroll == "Yes",  1, 0))
    ) %>%
    dplyr::select(-c(citizenship,restype))
      }

texas_applications <- dummy_vars(texas_applications)


texas_applications$satR <- as.numeric(texas_applications$satR)
texas_applications$actR <- as.numeric(texas_applications$actR)
texas_applications$testscoreR <- as.numeric(texas_applications$testscoreR)
texas_applications$gradyear <- as.numeric(texas_applications$gradyear)
```

Chunk 7

```
{r}
DT::datatable(
      texas_applications[1:25,],
      extensions = c('Scroller'),
      options = list(scrollY = 350,
                     scrollX = 500,
                     deferRender = TRUE,
                     scroller = TRUE,
                     dom = 'lBfrtip',
                     fixedColumns = TRUE,
                     searching = FALSE),
```

```
    rownames = FALSE)
```

Chunk 8

```
texas_applications %>%
  skim() %>%
  dplyr::select(skim_variable, complete_rate, n_missing,
                numeric.p0, numeric.p100) %>%
  dplyr::rename(variable=skim_variable, min=numeric.p0, max=numeric.p100) %>%
  mutate(complete_rate=round(complete_rate,2),
         min=round(min,2), max=round(max,2)) %>%
  arrange(variable) %>%
  nice_table()
```

Chunk 9

```
{r}
#remove data with NA more than 50%
texas_applications <- texas_applications %>%
dplyr::select(-c(actR,gradyear,hscentury,hseconstatus,hslos))
```

Chunk 10

```
{r}
texas_applications %>%
  count(admit) %>%
  mutate(perc = n *100/ nrow(texas_applications)) -> adas

ggplot(adas, aes(x = admit, y = perc)) +
  geom_bar(stat = "identity", colour='black', size=0.2)+
  xlab("If a student was admitted") +
  ylab("Percentage, %")
```

Chunk 11

```
{r discrete_plot}
texas_applications[,c("admit","male", "US_Citizen",
"Texas_resident","admit_prov","enroll")] %>%
  gather("variable","value") %>%
  group_by(variable) %>%
  count(value) %>%
  mutate(value = factor(value,levels=2:0)) %>%
```

```
  mutate(percent = n*100/134020) %>%
  ggplot(.,
  aes(variable,percent)) +
  geom_bar(stat = "identity", aes(fill = value)) +
  xlab("Variable") +
  ylab("Percentage") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_fill_manual(values = rev(c("#003f5c","#58508d", "#bc5090")))
```

Chunk 12

```
{r}
texas_applications %>%
  count(quartile) %>%
  mutate(perc = n *100/ nrow(texas_applications)) -> a1

texas_applications %>%
  count(decileR) %>%
  mutate(perc = n *100/ nrow(texas_applications)) -> b1

a <- ggplot(a1, aes(x = quartile, y = perc)) +
    geom_bar(stat = "identity", binwidth = 1, fill = "skyblue", color = "black", size=0.2)+
  labs(x = "Quartile", y = "Count")+ theme(axis.text.x = element_text(angle = 90, vjust =
0.5, hjust=1))

b <- ggplot(b1, aes(x = decileR, y = perc)) +
  geom_bar(stat = "identity", binwidth = 1, fill = "skyblue", color = "black", size=0.2)+
  labs(x = "Decile", y = "Percent") + theme(axis.text.x = element_text(angle = 90, vjust =
0.5, hjust=1))

plot<- ggarrange(a, b, ncol=2, nrow=1)

annotate_figure(plot, top = text_grob("Distribution of Student's HS class rank",
               color = "black", face = "bold", size = 14))
```

(Transform Data)

Chunk 13

```
{r message=FALSE, warning=FALSE,}

levels(texas_applications$hsprivate) <- c(levels(texas_applications$hsprivate),"None")
texas_applications$hsprivate[is.na(texas_applications$hsprivate)] <- "None"
```

```
levels(texas_applications$hsinstate) <- c(levels(texas_applications$hsinstate),"None")
texas_applications$hsinstate[is.na(texas_applications$hsinstate)] <- "None"

#texas_applications <- texas_applications %>% mutate(ethnic = ifelse(is.na(ethnic), "5",
ethnic))
#texas_applications$ethnic[is.na(texas_applications$ethnic)] <- "White, Non-Hispanic"
#texas_applications$ethnic <- as.factor(texas_applications$ethnic)

levels(texas_applications$ethnic) <- c(levels(texas_applications$ethnic),"None")
texas_applications$ethnic[is.na(texas_applications$ethnic)] <- "None"
texas_applications$ethnic <- combineLevels(texas_applications$ethnic,
                                    levs = c("White, Non-Hispanic","None"),
                                    newLabel = "White, Non-Hispanic")

# set values, dummy variable if top 10% or not
texas_applications <- texas_applications %>% mutate(top10 = ifelse(decileR == "Top
10%", TRUE, FALSE))
#texas_applications$top10 <- as.factor(texas_applications$top10)

levels(texas_applications$decileR) <- c(levels(texas_applications$decileR),"None")
texas_applications$decileR[is.na(texas_applications$decileR)] <- "None"


levels(texas_applications$quartile) <- c(levels(texas_applications$quartile),"None")
texas_applications$decileR[is.na(texas_applications$decileR)] <- "None"


#remove 52 observations where male=NA
texas_applications <- texas_applications %>%
  filter(!is.na(male))
```

Chunk 14

```
{r message=FALSE, warning=FALSE,}
# filter
attr_str <- c('admit', 'termdes',
'male','ethnic','US_Citizen','Texas_resident','satR','testscoreR',
            'top10','hsprivate','hsinstate','yeardes', 'enroll')

am_df <- texas_applications %>% dplyr::select(attr_str)
```

Chunk 15

```
{r message=FALSE, warning=FALSE,}

am_df %>%
  skim() %>%
  dplyr::select(skim_variable, complete_rate, n_missing,
                numeric.p0, numeric.p100) %>%
  dplyr::rename(variable=skim_variable, min=numeric.p0,
                max=numeric.p100) %>%
  mutate(complete_rate=round(complete_rate,2),
         min=round(min,2), max=round(max,2)) %>%
  arrange(variable) %>%
  nice_table()
```

Chunk 16

```
{r message=FALSE, warning=FALSE,}
#m_df <- merge_df %>% slice_sample(n=2000) %>%
m_df <- am_df %>%
  dplyr::select(where(is.numeric) & !c('admit','enroll')) %>%
  pivot_longer(!c('satR'), names_to='variable' , values_to = 'value') %>%
  drop_na()

m_df %>% ggplot(aes(x=value)) +
#m_df %>% ggplot(aes(x=value, group=avg_gpa, fill=avg_gpa)) +
geom_density(color='#023020') + facet_wrap(~variable, scales = 'free',  ncol = 4) +
theme_bw()
```

Chunk 17

```
{r message=FALSE, warning=FALSE,}
m_df <- am_df %>%
  dplyr::select((where(is.numeric) | c('admit','enroll')) & !c(yeardes)) %>%
  pivot_longer(!c('admit','enroll'), names_to='variable' , values_to = 'value') %>%
  drop_na()

m_df %>% ggplot(aes(y=value, x=admit, fill=enroll)) +
#m_df %>% ggplot(aes(x=value, group=TARGET_FLAG, fill=TARGET_FLAG)) +
geom_boxplot(color='#023020') + facet_wrap(~variable, scales = 'free',  ncol = 4) +
theme_bw()
```

Chunk 18

```
{r message=FALSE, warning=FALSE,}

#m_df <- am_df

m_df <- am_df %>%
  group_by(admit, ethnic, male) %>%
  summarise(n = n())

#treemap(m_df, index=c("ethnic","admit"), vSize="n", type="index")
treemap(m_df, index=c("admit","ethnic"), vSize="n", type="index",
      title="My Treemap",                    # Customize your title
      fontsize.title=12,
      align.labels=list(
        c("center", "center"),
        c("right", "bottom")
        ),                                 # Where to place labels in the rectangle?
    overlap.labels=0.5,                    #
    inflate.labels=F, )
```

Chunk 19

```
{r message=FALSE, warning=FALSE,}

m_df <- am_df %>%
  group_by(admit, ethnic, male) %>%
  summarise(n = n())

# plot
ggplot(m_df, aes(fill=ethnic, y=n, x=admit)) +
  geom_bar(position="stack", stat="identity") +
  scale_fill_viridis(discrete=TRUE, name="") +
  facet_wrap(~admit, scales = 'free',  ncol = 4)
  theme_ipsum() +
  ylab("Money input") +
  xlab("Month")
```

(Models)

Chunk 20

```
{r}
#no termdes=Fall or summer II only
pre_df <- am_df %>% dplyr::filter(yeardes < 1998)
b <- am_df %>% dplyr::filter(yeardes == 1998 & termdes=="Summer II")

pre_df <- rbind(pre_df,b)


demographic_df <- pre_df %>%
  dplyr::filter(yeardes == 1997) %>%
  group_by(admit,ethnic) %>%
  summarise(pre = n())
demographic_df$admit <- as.numeric(demographic_df$admit) -1

names(demographic_df) <- c('admit','ethnic','1997')
```

Chunk 21

```
{r}
pre_df %>%
  #dplyr::select(!c(studentid)) %>%
  skim() %>%
  dplyr::select(skim_variable, complete_rate, n_missing,
                numeric.p0, numeric.p100) %>%
  rename(variable=skim_variable, min=numeric.p0, max=numeric.p100) %>%
  mutate(complete_rate=round(complete_rate,2),
         min=round(min,2), max=round(max,2)) %>%
  arrange(variable) %>%
  nice_table()
```

Chunk 222

```
{r}
am_model_1 <- glm(pre_df, formula = admit ~ . -enroll -yeardes -termdes, family =
binomial(link = "probit"))
summary(am_model_1)
```

Chunk 23

```
{r}
am_model_1_aic <- am_model_1 %>% stepAIC(trace = FALSE)
summ(am_model_1_aic)
```

Chunk 24

```
{r}
a <- as.data.frame(summary(am_model_1_aic)$coefficients[,1])
a <- cbind(coef = rownames(a), a)
rownames(a) <- NULL
names(a) <- c('coef','< 1998')
a$model <- 'glm'
a <- a[, c("coef", "model", "< 1998")]

coef_tbl <- a
```

Chunk 25

```
{r}
coef_tbl <- glm_coef(am_df, '1998', coef_tbl)
coef_tbl <- glm_coef(am_df, '1999', coef_tbl)
coef_tbl <- glm_coef(am_df, '2000', coef_tbl)
coef_tbl <- glm_coef(am_df, '2001', coef_tbl)
coef_tbl <- glm_coef(am_df, '2002', coef_tbl)
```

Chunk 26

```
{r check_lm2}
par(mfrow=c(2,2))
plot(am_model_1_aic)
```

Chunk 27

```
{r model_2_linearity}
probabilities <- predict(am_model_1_aic, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")
head(predicted.classes)

#Only numeric predictors
data <- pre_df %>%
  dplyr::select_if(is.numeric)
```

```
predictors <- colnames(data)

# Bind the logit and tidying the data for plot
data <- data %>%
  mutate(logit = log(probabilities/(1-probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
```

Chunk 28

```
{r model_2_residuals}
am_model_1_aic.data <- augment(am_model_1_aic) %>%
  mutate(index = 1:n())

ggplot(am_model_1_aic.data, aes(index, .std.resid)) +
  geom_point(aes(color = admit), alpha = .5) +
  theme_bw()
```

Chunk 29

```
{r warning=FALSE, message=FALSE}
car::mmps(am_model_1_aic, span = 3/4, layout = c(2, 2))
```

Chunk 30

```
{r model_2_vif}
car::vif(am_model_1_aic)
```

(Predict)

Chunk 31

```
{r}
demographic_df <- demographicCount(am_df, am_model_1_aic, 1998, demographic_df)
demographic_df <- demographicCount(am_df, am_model_1_aic, 1999, demographic_df)
demographic_df <- demographicCount(am_df, am_model_1_aic, 2000, demographic_df)
demographic_df <- demographicCount(am_df, am_model_1_aic, 2001, demographic_df)
demographic_df <- demographicCount(am_df, am_model_1_aic, 2002, demographic_df)
```

(Lasso)

Chunk 32

```
{r}
# set seed for consistancy
set.seed(42)

# build X matrix and Y vector
X <- model.matrix(admit ~ . -enroll -yeardes -termdes , data=pre_df)[,-1]
Y <- pre_df["admit"]

demographic_lasso_df <- pre_df %>%
  dplyr::filter(yeardes == 1997) %>%
  group_by(admit,ethnic) %>%
  summarise(pre = n())

demographic_lasso_df$admit <- as.numeric(demographic_lasso_df$admit) - 1

names(demographic_lasso_df) <- c('admit','ethnic','1997')
```

Chunk 33

```
{r}
lasso.model<- cv.glmnet(x=X,y=Y,
                        family = "binomial",
                        link = "probit",
                        standardize = TRUE,                    #standardize
                        nfold = 5,
                        alpha=1)                               #alpha=1 is lasso

l.min <- lasso.model$lambda.min
l.1se <- lasso.model$lambda.1se
coef(lasso.model, s = "lambda.min" )
coef(lasso.model, s = "lambda.1se" )
lasso.model
```

Chunk 34

```
{r}
a <- coeff2dt(fitobject = lasso.model, s = "lambda.min")
names(a) <- c('coef','< 1998')
a$model <- 'lasso'
a <- a[, c("coef", "model", "< 1998")]

coef_lasso_tbl <- a
```

Chunk 35

```
{r}
coef_lasso_tbl <- lasso_coef(am_df, 1998 ,  coef_lasso_tbl)
coef_lasso_tbl <- lasso_coef(am_df, 1999 ,  coef_lasso_tbl)
coef_lasso_tbl <- lasso_coef(am_df, 2000 ,  coef_lasso_tbl)
coef_lasso_tbl <- lasso_coef(am_df, 2001 ,  coef_lasso_tbl)
coef_lasso_tbl <- lasso_coef(am_df, 2002 ,  coef_lasso_tbl)
```

Chunk 36

```
{r}
par(mfrow=c(2,2))

plot(lasso.model)
plot(lasso.model$glmnet.fit, xvar="lambda", label=TRUE)
plot(lasso.model$glmnet.fit, xvar='dev', label=TRUE)

rocs <- roc.glmnet(lasso.model, newx = X, newy = Y )
plot(rocs,type="l")
```

Chunk 37

```
{r}
assess.glmnet(lasso.model,
              newx = X,
              newy = Y )

print(glmnet_cv_aicc(lasso.model, 'lambda.min'))
print(glmnet_cv_aicc(lasso.model, 'lambda.1se'))
```

Chunk 38

```
{r}
as.data.frame(as.matrix(coef(lasso.model, s = "lambda.min"))) %>%
  arrange(desc(s1)) %>%
  nice_table(cap='Model Coefficients', cols=c('Est'))
```

Chunk 39

```
{r}
vip(lasso.model, num_features=20 ,geom = "col", include_type=TRUE, lambda =
"lambda.min")
```

```
coeffs.table <- coeff2dt(fitobject = lasso.model, s = "lambda.min")

coeffs.table %>% mutate(name = fct_reorder(name, desc(coefficient))) %>%
ggplot() +
  geom_col(aes(y = name, x = coefficient, fill = {coefficient > 0})) +
  xlab(label = "") +
  ggtitle(expression(paste("Lasso Coefficients with ", lambda, " = 0.0275"))) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust=0.5),legend.position =
"none")
```

(Model Results)

Chunk 40

```
{r}
# Create matrix new data
am_lasso1998_df <- am_df %>% dplyr::filter(yeardes == 1998 & termdes=="Fall") %>%
  dplyr::select(!c(enroll,yeardes,termdes))

X_test <- model.matrix(admit ~ . ,data=am_lasso1998_df)[,-1]
Y_test <- am_lasso1998_df["admit"]

# predict using coefficients at lambda.min
lassoPred <- predict(lasso.model, newx = X_test, type = "response", s = 'lambda.min')

#pred_df <- am_lasso1998_df
am_lasso1998_df$admit_prob <- lassoPred[,1]
am_lasso1998_df$admit_pred <- ifelse(lassoPred >= 0.5, 1, 0)[,1]
```

Chunk 41

```
{r eval=FALSE}

confusion.glmnet(lasso.model, newx = X_test, newy = Y_test, s = 'lambda.min')
```

Chunk 42

```
{r}
demographic_lasso_df <- demographicCountLasso(am_df, lasso.model, 1998,
demographic_lasso_df)
demographic_lasso_df <- demographicCountLasso(am_df, lasso.model, 1999,
```

```
demographic_lasso_df)
demographic_lasso_df <- demographicCountLasso(am_df, lasso.model, 2000,
demographic_lasso_df)
demographic_lasso_df <- demographicCountLasso(am_df, lasso.model, 2001,
demographic_lasso_df)
demographic_lasso_df <- demographicCountLasso(am_df, lasso.model, 2002,
demographic_lasso_df)
```

(Analysis)

Chunk 43

```
{r}
m_df <- am_df %>%
  dplyr::filter(admit == 1) %>%
  dplyr::select(admit, ethnic, yeardes) %>%
  group_by(ethnic,yeardes) %>%
  summarise(count = n())

d <- transform(m_df, percent = ave(count,yeardes,FUN = prop.table))

d %>%
    ggplot(aes(x=yeardes, y = percent, colour = ethnic, group = ethnic)) +
    geom_line() +
    geom_point(size=1) +
    labs(title='Admission Demographics Texas A&M Actual') +
    scale_x_continuous(n.breaks=11)
    theme_bw()


# display table
d <- d %>% dplyr::select(!c(count)) %>%
  pivot_wider(names_from = yeardes, values_from = percent ,names_sort = TRUE)

d %>% nice_table(cap='Admission Demographics Actual')
```

Chunk 44

```
{r}
a_df <- demographic_df %>%
```

```
  dplyr::filter(admit == 1) %>%
  #dplyr::select(admit, ethnic, yeardes) %>%
  pivot_longer(!c('ethnic','admit'), names_to='variable' , values_to = 'value')

a <- transform(a_df, percent = ave(value,variable,FUN = prop.table))
a$variable = as.integer(a$variable)

m_df <- am_df %>% dplyr::filter(yeardes > 1997) %>%
  dplyr::filter(admit == 1) %>%
  dplyr::select(admit, ethnic, yeardes) %>%
  group_by(ethnic,yeardes) %>%
  summarise(count = n())

b <- transform(m_df, percent = ave(count,yeardes,FUN = prop.table))
names(b) <- c('ethnic','variable','actual_count','actual_per')
b$variable = as.integer(b$variable)

c <- a %>% full_join(b, by=c('ethnic','variable'))

c$dif_per <- c$percent - c$actual_per

c %>%
    ggplot(aes(x=variable, y=dif_per, colour = ethnic, group = ethnic)) +
    geom_line() +
    geom_point(size=1) +
    labs(title='Admission Demographics (predicted - actual)') +
    xlab("Year") +
    ylab("Difference: Predicted-Actual") +
    theme_bw()

c %>% dplyr::select(ethnic,variable,dif_per) %>% drop_na() %>%
  pivot_wider(names_from = variable, values_from = dif_per ,names_sort = TRUE) %>%
  nice_table(cap='Actual vs Predicted Lasso')
```

Chunk 45

```
{r}
a_df <- demographic_lasso_df %>%
  dplyr::filter(admit == 1) %>%
  #dplyr::select(admit, ethnic, yeardes) %>%
  pivot_longer(!c('ethnic','admit'), names_to='variable' , values_to = 'value')

a <- transform(a_df, percent = ave(value,variable,FUN = prop.table))
a$variable = as.integer(a$variable)
```

```
m_df <- am_df %>% dplyr::filter(yeardes > 1996) %>%
  dplyr::filter(admit == 1) %>%
  dplyr::select(admit, ethnic, yeardes) %>%
  group_by(ethnic,yeardes) %>%
  summarise(count = n())

b <- transform(m_df, percent = ave(count,yeardes,FUN = prop.table))
names(b) <- c('ethnic','variable','actual_count','actual_per')
b$variable = as.integer(b$variable)

c <- a %>% full_join(b, by=c('ethnic','variable'))

c$dif_per <- c$percent - c$actual_per

c %>% dplyr::filter(variable > 1997) %>%
    ggplot(aes(x=variable, y=dif_per, colour = ethnic, group = ethnic)) +
    geom_line() +
    geom_point(size=1) +
    labs(title='Admission Demographics (predicted - actual)') +
    xlab("Year") +
    ylab("Difference: Predicted-Actual") +
    theme_bw()

c %>% dplyr::filter(variable > 1997) %>%
  dplyr::select(ethnic,variable,dif_per) %>% drop_na() %>%
  pivot_wider(names_from = variable, values_from = dif_per ,names_sort = TRUE) %>%
  nice_table(cap='Admission Demographics (predicted - actual)')
```

(GLM)

Chunk 46

```
{r}

m_df <- coef_tbl %>%
  #dplyr::select(where(is.numeric) & !c('admit','enroll')) %>%
  dplyr::filter(coef != '(Intercept)') %>%
  pivot_longer(!c('coef','model'), names_to='variable' , values_to = 'value') %>%
  drop_na()

ggplot(data=m_df, aes(x=variable, y = value, colour = coef, group = coef)) +
  geom_line() +
  geom_point(size=0.5) +
  facet_wrap( ~coef, ncol = 4, as.table=TRUE, labeller = "label_both") +
```

```
  labs(title='glm() Admission Coefficients') +
  theme_bw()

m_df %>%
    ggplot(aes(x=variable, y=value, colour = coef, group = coef)) +
    geom_line() +
    geom_point(size=1) +
    labs(title='glm() Admission Coefficients') +
    theme_bw()

coef_tbl %>% dplyr::select(!c(model)) %>% nice_table(cap='glm() Admission
Coefficients')
```

Chunk 47

```
{r}
m_df <- coef_lasso_tbl %>%
  #dplyr::select(where(is.numeric) & !c('admit','enroll')) %>%
  dplyr::filter(coef != '(Intercept)') %>%
  pivot_longer(!c('coef','model'), names_to='variable' , values_to = 'value') %>%
  drop_na()

ggplot(data=m_df, aes(x=variable, y = value, colour = coef, group = coef)) +
  geom_line() +
  geom_point(size=0.5) +
  facet_wrap( ~coef, ncol = 4, as.table=TRUE, labeller = "label_both") +
  labs(title='Lasso Admission Coefficients') +
  theme_bw()

m_df %>%
    ggplot(aes(x=variable, y=value, colour = coef, group = coef)) +
    geom_line() +
    geom_point(size=1) +
    labs(title='Lasso Admission Coefficients') +
    theme_bw()

coef_lasso_tbl %>% dplyr::select(!c(model)) %>% nice_table(cap='Lasso Admission
Coefficients')
```

Chunk 48

```
{r}
d <- demographic_df %>% dplyr::filter(admit == 1) %>%
        pivot_longer(!c('admit','ethnic'), names_to='variable' , values_to = 'value')

d <- transform(d, percent = ave(value,variable,FUN = prop.table))

d %>%
    ggplot(aes(x=variable, y = percent, colour = ethnic, group = ethnic)) +
    geom_line() +
    geom_point(size=0.5) +
    labs(title='glm() Admission Demographics') +
    facet_wrap( ~ethnic, ncol = 4, as.table=TRUE, labeller = "label_both", scales="free_y")
+
    theme_bw()

d %>%
    ggplot(aes(x=variable, y = percent, colour = ethnic, group = ethnic)) +
    geom_line() +
    labs(title='glm() Admission Demographics') +
    geom_point(size=1) +
    theme_bw()

# display table
d <- d %>% dplyr::select(!c(value,admit)) %>%
  pivot_wider(names_from = variable, values_from = percent ,names_sort = TRUE)

d %>% nice_table(cap='glm() Admission Demographics')
```

Chunk 49

```
{r}
d <- demographic_lasso_df %>% dplyr::filter(admit == 1) %>%
        pivot_longer(!c('admit','ethnic'), names_to='variable' , values_to = 'value')

d <- transform(d, percent = ave(value,variable,FUN = prop.table))

d %>%
    ggplot(aes(x=variable, y = percent, colour = ethnic, group = ethnic)) +
    geom_line() +
    geom_point(size=0.5) +
    labs(title='Lasso Admission Demographics') +
    facet_wrap( ~ethnic, ncol = 4, as.table=TRUE, labeller = "label_both", scales="free_y")
+
    theme_bw()
```

```
d %>%
    ggplot(aes(x=variable, y = percent, colour = ethnic, group = ethnic)) +
    geom_line() +
    labs(title='Lasso Admission Demographics') +
    geom_point(size=1) +
    theme_bw()

# display table
d <- d %>% dplyr::select(!c(value,admit)) %>%
  pivot_wider(names_from = variable, values_from = percent ,names_sort = TRUE)

d %>% nice_table(cap='Lasso Admission Demographics')
```