An Empirical Study of Factors Influencing Application Volume at U.S. Postsecondary Institutions

William J. Hassel

Abstract

This project investigates factors influencing the number of applications received by liberal arts colleges in the United States, focusing on how public-facing variables such as test scores, tuition costs, and financial aid interact with institutional characteristics. Using data from the 2022–2023 admissions cycle, I estimated a multiple regression model where the log of applicant count is regressed on acceptance rate, retention rate, standardized test scores, cost of attendance, and percentage of students receiving financial aid, with several interaction and nonlinear terms included. Results suggest that the effect of acceptance rate, cost, and percent aid on application volume depends on a college's retention rate, and that test-optional schools receive significantly more applicants. These findings illustrate how colleges' appeal to prospective students may be shaped by the combined presentation of several key metrics.

Introduction

Prior research has shown that variables such as acceptance rate, standardized test scores, and financial aid offerings are associated with application volume (Hossler et al., 1999). However, many of these factors do not operate independently. For instance, the perceived desirability of a low acceptance rate may vary depending on other signals of institutional quality, such as student retention. Similarly, the influence of financial aid availability may differ depending on a college's overall cost of attendance.

These interdependencies suggest that prospective students may be responding to a more complex set of institutional signals than any single variable can capture. Colleges do not present themselves to applicants in isolation, so information about selectivity, cost, aid, and student success are often evaluated together, both in public datasets and in rankings or promotional materials. As a result, understanding how these characteristics interact may be key to explaining variation in application behavior.

My goal is to explore how combinations of commonly reported variables affect application behavior using admissions data from the 2022–2023 cycle, aiming to contribute a more nuanced understanding of how public-facing institutional data may shape student decision-making. This analysis focuses on four-year institutions and emphasizes interaction and nonlinear effects that may otherwise be overlooked in simpler additive models.

Data

All data used in this project was gathered from the Integrated Postsecondary Education Data System (IPEDS), a service of the United States Department of Education that conducts annual surveys of every postsecondary institution in the country.

The cases for this dataset are the 1200 most applied to four-year postsecondary institutions in the United States. The response variable, how many applications each institution received in 2023. The predictor variables are acceptance rate, cost of tuition and fees, retention rate, percentage of students that received financial aid, and whether or not a school is test-optional (all collected from the 2022-2023 academic year).

Results

To analyze the impact of each predictor I assumed an ordinary linear model

$$\log Y = \beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 \log X_4+\beta_5 X_5+\beta_6 X_5^2+\beta_7 X_2 X_3+\beta_8 X_2 X_4+\beta_9 X_2 X_5+\epsilon, \epsilon \overset{\text{i.i.d.}}{\sim} N(0,\sigma).$$

When fitted to the data the model gave the following coefficient and standard error estimates, all of which resulted in significant p-values.

Coefficient	Associated Predictor	Estimate	SE Estimate
β_0	Intercept	-15.1	3.67
β_1	`testScores`	-0.303	0.109
β_2	`retentionRt`	0.287	4.13e-02
β_3	`acceptRt`	3.80e-02	1.11e-02
β_4	`log(cost)`	0.646	0.290
β_5	`percentAid`	0.208	4.55e-02
β_6	`percentAid`^{2}	-9.72e-04	1.81e-04
β_7	`retentionRt*acceptRt`	-5.19e-04	1.40e-04
β_8	`retentionRt*log(cost)`	-1.31e-02	3.72e-03
β_9	`retentionRt*percentAid`	-7.60e-04	2.98e-04

Note that the response variable, `applicants2023`, and the predictor `cost` are log transformed. The predictor `percentAid` also has an additional quadratic term. There are also three significant interaction term in this model.

The first of these notable features is the transformation of the response variable, which I did to account for the heavy non-linearity between applicants in 2023 and each predictor variable. A significant example of this is shown in Fig. 1.

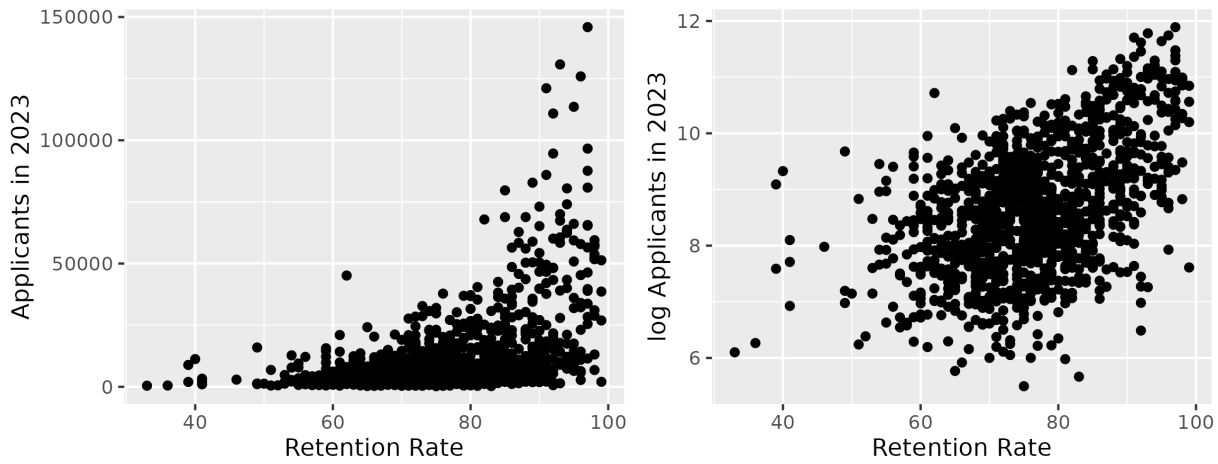


Figure 1: Relationship between retention rate and both unaltered and log transformed volume of applicants in 2023.

The transformations on the two predictors originated from analyzing residual plots and finding extreme non-linearity in `cost` and non-constant variance in `percentAid`. These transformations are to be expected, as cost varies from very low to very high values, and percent aid follows a roughly quadratic shape as shown in Fig. 2.

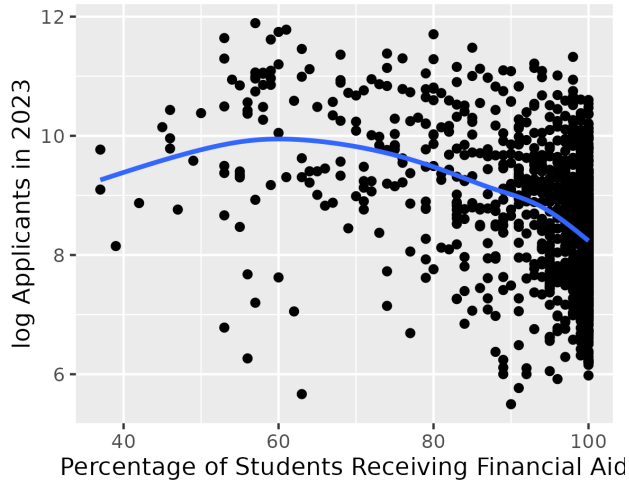


Figure 2: Approximately quadratic relationship between percent of students receiving aid and applicant volume in 2023.

After applying these transformations I built a total of ten models, each with a single possible interaction term between two predictors, and found that when compared to the original reduced model via a LRT test, the only significant terms were between `retentionRt` and each of `acceptRt` (p-value: 1.21e-06), `log(cost)` (p-value: 0.010), and `percentAid` (p-value: 6.80e-04). An example of one of these interactions is illustrated visually in Fig. 3, as different levels of retention rate drastically impact the relationship between acceptance rate and applicants in 2023.

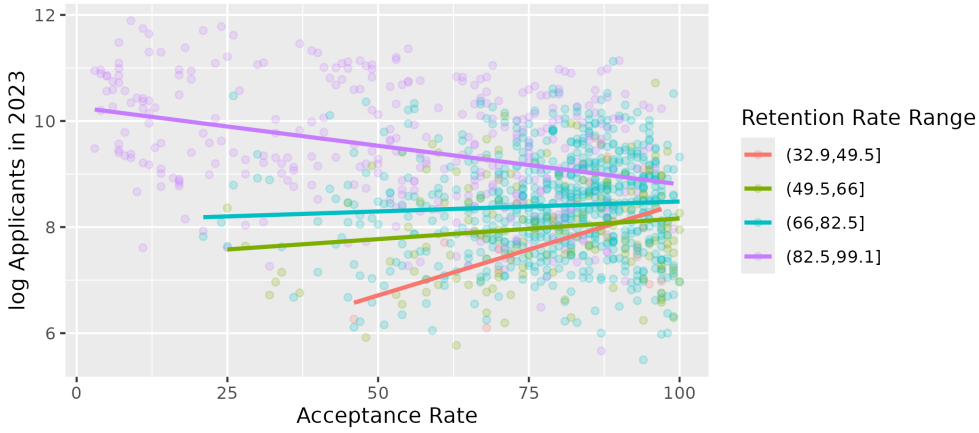


Figure 3: Relationship between acceptance rate and volume of applicants in 2023, seperated by ranges of retention rate.

Diagnostic plots are shown in Appendix A.

Discussion

Interpreting Results

This model presents several interesting insights into the college admission process, particularly when looking at the significant interaction terms. Assume all other covariates are fixed for each interpretation.

For the first term, which is displayed in Fig. 3, the model predicts that an increase of a single percentage of acceptance rate increases 2023 applicant volume by an average of 2.1% at the lowest retention rate of 32.9%. At the median retention rate of 76.5%, the same increase in acceptance rate decreases 2023 applicant volume by 0.2% on average, and at the highest retention rate of 99.1%, it decreases the applicant volume by 1.3% on average. This trend is fascinating, as it tells us that at institutions with very low retention rates, a school that accepts a larger proportion of students will get drastically more applicants. Conversely, at schools with very high retention rates, the more selective they are the more applicants they will get on average.

For second interaction term, between `retentionRt` and `log(cost)`, the model predicts that doubling the cost results in an a mean increase of 16.1% in 2023 applicant volume for the lowest retention rate of 32.9%. In contrast, for the median retention rate, doubling the cost decreases applicant volume by 21.9% on average , and for the highest retention rate, by 36.4%.

Next, for the interaction between `retentionRt` and `percentAid`, the latter has a concave quadratic relationship with the response variable. The model predicts that at the minimum retention rate, the maximum applicants, or turning point of the curve, is when 94.1% of students receive financial aid, on average. At the median retention rate this drops to a mean of 77.1% of students, and at the maximum retention rate 68.3% of students.

Finally, note that when compared to test optional schools, the model predicts that requiring students to submit test scores results in a mean decrease of 26.1% in 2023 applicant volume.

Ethical Concerns

One notable limitation of this analysis is lack of representation of smaller, less funded colleges in the dataset. When cleaning the IDEPS data, I found that schools with less that 300 students more often then not had incomplete variable reporting in the categories I was looking for. This rendered them impractical for the purposes of this project, even though they are still relevant members of the population and likely contained important information for linear regression analysis.

Although the results from the model are still relevant, this lack of completeness in the data can still bring up ethical concerns. The 2025 NeurIPS Code of Ethics advises against models or datasets that "encode, contain or exacerbate bias," which may be present here, especially in some of the interaction terms that condition by retention rate.

References

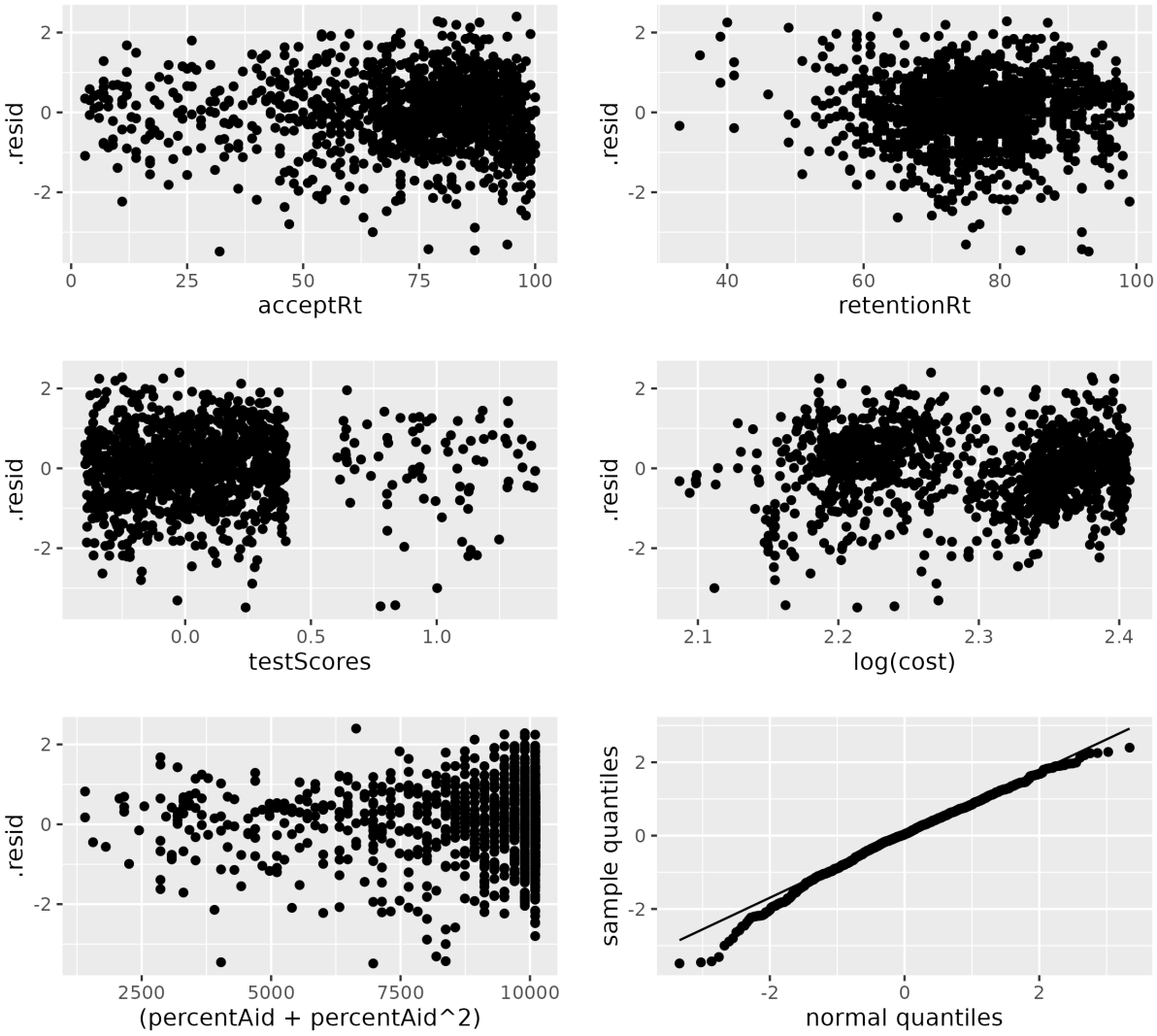
Hossler, D., Schmit, J., & Vesper, N. (1999). *Going to College: How Social, Economic, and Educational Factors Influence the Decisions Students Make.* Johns Hopkins University Press.

U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), 2022-2023.

NeurIPS Code of Ethics. (2025). Nips.cc; NeurIPS. https://nips.cc/public/EthicsGuidelines

Appendices

Appendix A: Diagnostic Plots



Appendix B: Raw R Code

```
---
output: pdf_document
title: "An Empirical Study of Factors Influencing Application Volume at U.S. Postsecondary
Institutions"
author: William J. Hassel
affiliation: Carleton College
abstract:
    "This project investigates factors influencing the number of applications received by
liberal arts colleges in the United States, focusing on how public-facing variables such as test
scores, tuition costs, and financial aid interact with institutional characteristics. Using data
from the 2022–2023 admissions cycle, I estimated a multiple regression model where the log of
applicant count is regressed on acceptance rate, retention rate, standardized test scores, cost
of attendance, and percentage of students receiving financial aid, with several interaction and
nonlinear terms included. Results suggest that the effect of acceptance rate, cost, and percent
aid on application volume depends on a college's retention rate, and that test-optional schools
receive significantly more applicants. These findings illustrate how colleges' appeal to
prospective students may be shaped by the combined presentation of several key metrics."
geometry: margin=1in
fontsize: 11pt
indent: false
header-includes:
  - \setlength\parindent{0pt}
  - \setlength\parskip{1em}
  - \usepackage{graphicx}
  - \usepackage{float}
---

```{r setup, include=FALSE}
library(ggplot2)
library(patchwork)
library(dplyr)
library(broom)
library(GGally)
library(knitr)
knitr::opts_chunk$set(echo = TRUE,
                      fig.align = "center",
                      fig.pos = "H")
```

Introduction

Prior research has shown that variables such as acceptance rate, standardized test scores, and
financial aid offerings are associated with application volume (Hossler et al., 1999). However,
many of these factors do not operate independently. For instance, the perceived desirability of
a low acceptance rate may vary depending on other signals of institutional quality, such as
student retention. Similarly, the influence of financial aid availability may differ depending
on a college's overall cost of attendance.

These interdependencies suggest that prospective students may be responding to a more complex
set of institutional signals than any single variable can capture. Colleges do not present
themselves to applicants in isolation, so information about selectivity, cost, aid, and student
success are often evaluated together, both in public datasets and in rankings or promotional
materials. As a result, understanding how these characteristics interact may be key to
explaining variation in application behavior.

My goal is to explore how combinations of commonly reported variables affect application
behavior using admissions data from the 2022–2023 cycle, aiming to contribute a more nuanced
understanding of how public-facing institutional data may shape student decision-making. This
analysis focuses on four-year institutions and emphasizes interaction and nonlinear effects that
may otherwise be overlooked in simpler additive models.

Data
```

```{r, echo=FALSE, include=FALSE}
# Importing and cleaning data

colleges2022 <- read.csv("2022colleges.csv")
colleges2023 <- read.csv("admissions2023.csv")

colleges <- subset(colleges2022, ADM2022_RV.Applicants.total != "")
colleges <- subset(colleges, ADM2022_RV.Applicants.total >= 300)
colleges <- merge(colleges, colleges2023, by="unitid")

colleges <- colleges[, c(1:10,13)]
colleges <- colleges %>%
  rename(name=institution.name.x, year=year.x,
         applicants2022=ADM2022_RV.Applicants.total,
         applicants2023=ADM2023.Applicants.total,
         acceptRt=DRVADM2022_RV.Percent.admitted...total,
         cost=DRVIC2022.Tuition.and.fees..2022.23,
         retentionRt=EF2022D_RV.Full.time.retention.rate..2022,

percentAid=SFA2122_RV.Percent.of.full.time.first.time.undergraduates.awarded.any.financial.aid,
         GPA=ADM2022_RV.Secondary.school.GPA,
         testScores=ADM2022_RV.Admission.test.scores)

colleges <- na.omit(colleges)

colleges$GPA <- ifelse(colleges$GPA ==
                         "Required to be considered for admission", 1, 0)
colleges$testScores <- ifelse(colleges$testScores ==
                         "Required to be considered for admission", 1, 0)
```

All data used in this project was gathered from the Integrated Postsecondary Education Data System (IPEDS), a service of the United States Department of Education that conducts annual surveys of every postsecondary institution in the country.

The cases for this dataset are the 1200 most applied to four-year postsecondary institutions in the United States. The response variable, how many applications each institution received in 2023. The predictor variables are acceptance rate, cost of tuition and fees, retention rate, percentage of students that received financial aid, and whether or not a school is test-optional (all collected from the 2022-2023 academic year).

Results

To analyze the impact of each predictor I assumed an ordinary linear model
$$\log Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4\log X_4+\beta_5X_5+\beta_6X_5^2+
\beta_7X_2X_3+\beta_8X_2X_4+\beta_9X_2X_5+\epsilon,\epsilon\overset{\mathrm{i.i.d.}}{\sim}N(0,
\sigma).$$
When fitted to the data the model gave the following coefficient and standard error estimates, all of which resulted in significant p-values.

```{r, echo=FALSE, include=FALSE}
colleges.lm <- lm(data=colleges, log(applicants2023) ~ testScores + retentionRt *
                    acceptRt + retentionRt * log(cost) + retentionRt *
                    percentAid + I(percentAid^2))

summary(colleges.lm)
```

Coefficient	Associated Predictor	Estimate	SE Estimate
β_0	Intercept	-15.1	3.67

β_1	`testScores`	-0.303	0.109
β_2	`retentionRt`	0.287	4.13e-02
β_3	`acceptRt`	3.80e-02	1.11e-02
β_4	`log(cost)`	0.646	0.290
β_5	`percentAid`	0.208	4.55e-02
β_6	`percentAid`^{2}	-9.72e-04	1.81e-04
β_7	`retentionRt*acceptRt`	-5.19e-04	1.40e-04
β_8	`retentionRt*log(cost)`	-1.31e-02	3.72e-03
β_9	`retentionRt*percentAid`	-7.60e-04	2.98e-04

Note that the response variable, `applicants2023`, and the predictor `cost` are log transformed. The predictor `percentAid` also has an additional quadratic term. There are also three significant interaction term in this model.

The first of these notable features is the transformation of the response variable, which I did to account for the heavy non-linearity between applicants in 2023 and each predictor variable. A significant example of this is shown in Fig. 1.

```{r, echo=FALSE, include=FALSE}
plot1 <- ggplot(data=colleges, aes(x=retentionRt, y=(applicants2023))) +
  geom_point() + labs(x="Retention Rate", y="Applicants in 2023")
plot2 <- ggplot(data=colleges, aes(x=retentionRt, y=log(applicants2023))) +
  geom_point() + labs(x="Retention Rate", y="log Applicants in 2023")

linearityPlot <- plot1 + plot2

ggsave("linearityPlot.png", linearityPlot, width = 20, height = 8, units = "cm", dpi = 300)
```

```{r, echo=FALSE, fig.cap="Relationship between retention rate and both unaltered and log transformed volume of applicants in 2023.", out.width="100%"}
knitr::include_graphics("linearityPlot.png")
```

The transformations on the two predictors originated from analyzing residual plots and finding extreme non-linearity in `cost` and non-constant variance in `percentAid`. These transformations are to be expected, as cost varies from very low to very high values, and percent aid follows a roughly quadratic shape as shown in Fig. 2.

```{r, echo=FALSE, include=FALSE}
plot3 <- ggplot(data=colleges, aes(x=percentAid, y=log(applicants2023))) +
  geom_point() + geom_smooth(se=FALSE) +
  labs(x="Percentage of Students Receiving Financial Aid", y="log Applicants in 2023")

ggsave("aidPlot.png", plot3, width = 10, height = 8, units = "cm", dpi = 300)
```

```{r, echo=FALSE, fig.cap="Approximately quadratic relationship between percent of students receiving aid and applicant volume in 2023.", out.width="50%"}
knitr::include_graphics("aidPlot.png")
```

```{r, echo=FALSE, include=FALSE}
collegesA.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt + retentionRt
                   + testScores + log(cost) + percentAid + I(percentAid^2))
colleges2.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt * testScores +
                     retentionRt + log(cost) + percentAid + I(percentAid^2))
colleges3.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt * log(cost) +
                     retentionRt + testScores + percentAid + I(percentAid^2))
colleges4.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt * percentAid +
                     retentionRt + testScores + log(cost) + I(percentAid^2))
colleges5.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt + retentionRt
                   * testScores + log(cost) + percentAid + I(percentAid^2))
colleges6.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt + retentionRt
```

```
                    * log(cost) + testScores + percentAid + I(percentAid^2))
colleges7.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt + retentionRt
                    * percentAid + testScores + log(cost) + I(percentAid^2))
colleges8.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt + retentionRt
                    + testScores * log(cost) + percentAid + I(percentAid^2))
colleges9.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt + retentionRt
                    + testScores * percentAid + log(cost) + I(percentAid^2))
colleges10.lm <- lm(data=colleges, log(applicants2023) ~ acceptRt + retentionRt
                    + testScores + log(cost) * percentAid + I(percentAid^2))

anova(collegesA.lm, colleges.lm, test="LRT")
anova(collegesA.lm, colleges2.lm, test="LRT")
anova(collegesA.lm, colleges3.lm, test="LRT")
anova(collegesA.lm, colleges4.lm, test="LRT")
anova(collegesA.lm, colleges5.lm, test="LRT")
anova(collegesA.lm, colleges6.lm, test="LRT")
anova(collegesA.lm, colleges7.lm, test="LRT")
anova(collegesA.lm, colleges8.lm, test="LRT")
anova(collegesA.lm, colleges9.lm, test="LRT")
anova(collegesA.lm, colleges10.lm, test="LRT")
```

After applying these transformations I built a total of ten models, each with a single possible interaction term between two predictors, and found that when compared to the original reduced model via a LRT test, the only significant terms were between `retentionRt` and each of `acceptRt` (p-value: 1.21e-06), `log(cost)` (p-value: 0.010), and `percentAid` (p-value: 6.80e-04). An example of one of these interactions is illustrated visually in Fig. 3, as different levels of retention rate drastically impact the relationship between acceptance rate and applicants in 2023.

```{r, echo=FALSE, include=FALSE}
colleges$ret_bin <- cut(colleges$retentionRt, breaks = 4)

plot4 <- ggplot(colleges, aes(x = acceptRt, y = log(applicants2023), color = ret_bin)) +
  geom_point(alpha=0.2) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x="Acceptance Rate", y="log Applicants in 2023",
       color="Retention Rate Range")

ggsave("interactionPlot.png", plot4, width = 18, height = 8, units = "cm", dpi = 300)
```

```{r, echo=FALSE, fig.cap="Relationship between acceptance rate and volume of applicants in 2023, seperated by ranges of retention rate.", out.width="80%"}
knitr::include_graphics("interactionPlot.png")
```

Diagnostic plots are shown in Appendix A.

Discussion

Interpreting Results

This model presents several interesting insights into the college admission process, particularly when looking at the significant interaction terms. Assume all other covariates are fixed for each interpretation.

For the first term, which is displayed in Fig. 3, the model predicts that an increase of a single percentage of acceptance rate increases 2023 applicant volume by an average of 2.1% at the lowest retention rate of 32.9%. At the median retention rate of 76.5%, the same increase in acceptance rate decreases 2023 applicant volume by 0.2% on average, and at the highest retention rate of 99.1%, it decreases the applicant volume by 1.3% on average. This trend is fascinating,

as it tells us that at institutions with very low retention rates, a school that accepts a larger proportion of students will get drastically more applicants. Conversely, at schools with very high retention rates, the more selective they are the more applicants they will get on average.

For second interaction term, between `retentionRt` and `log(cost)`, the model predicts that doubling the cost results in an a mean increase of 16.1% in 2023 applicant volume for the lowest retention rate of 32.9%. In contrast, for the median retention rate, doubling the cost decreases applicant volume by 21.9% on average , and for the highest retention rate, by 36.4%.

Next, for the interaction between `retentionRt` and `percentAid`, the latter has a concave quadratic relationship with the response variable. The model predicts that at the minimum retention rate, the maximum applicants, or turning point of the curve, is when 94.1% of students receive financial aid, on average. At the median retention rate this drops to a mean of 77.1% of students, and at the maximum retention rate 68.3% of students.

Finally, note that when compared to test optional schools, the model predicts that requiring students to submit test scores results in a mean decrease of 26.1% in 2023 applicant volume.

Ethical Concerns

One notable limitation of this analysis is lack of representation of smaller, less funded colleges in the dataset. When cleaning the IDEPS data, I found that schools with less that 300 students more often then not had incomplete variable reporting in the categories I was looking for. This rendered them impractical for the purposes of this project, even though they are still relevant members of the population and likely contained important information for linear regression analysis.

Although the results from the model are still relevant, this lack of completeness in the data can still bring up ethical concerns. The 2025 NeurIPS Code of Ethics advises against models or datasets that "encode, contain or exacerbate bias," which may be present here, especially in some of the interaction terms that condition by retention rate.

\newpage

References

\hangindent=0.7cm Hossler, D., Schmit, J., & Vesper, N. (1999). *Going to College: How Social, Economic, and Educational Factors Influence the Decisions Students Make*. Johns Hopkins University Press.

\hangindent=0.7cm U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), 2022-2023.

\hangindent=0.7cm *NeurIPS Code of Ethics*. (2025). Nips.cc; NeurIPS. https://nips.cc/public/EthicsGuidelines

Appendices

Appendix A: Diagnostic Plots

```{r, echo=FALSE, include=FALSE}
aug <- augment(colleges.lm)
aug <- aug %>%
  rename(cost = "log(cost)")

# Apply some margin to prevent clipping
common_theme <- theme(plot.margin = margin(10, 10, 10, 10))

plot1 <- ggplot(aug, aes(y = .resid, x = acceptRt)) + geom_point() + common_theme
plot2 <- ggplot(aug, aes(y = .resid, x = retentionRt)) + geom_point() + common_theme
plot3 <- ggplot(aug, aes(y = .resid, x = testScores)) + geom_jitter() + common_theme
```

```
plot4 <- ggplot(aug, aes(y = .resid, x = log(cost))) + geom_point() + common_theme
plot5 <- ggplot(aug, aes(y = .resid, x = (percentAid + percentAid^2))) + geom_point() +
common_theme
plot6 <- ggplot(aug, aes(sample = .resid)) + geom_qq() +
  geom_qq_line() + labs(y = "sample quantiles", x = "normal quantiles") +
  common_theme

# Combine plots without specifying height here
final_plot <- (plot1 + plot2) /
              (plot3 + plot4) /
              (plot5 + plot6)

# Save to file with exact height: 3 rows * 4 cm = 12 cm
ggsave("residuals_plot.png", final_plot, width = 20, height = 18, units = "cm", dpi = 300)
```

```{r, echo=FALSE, out.width="100%"}
knitr::include_graphics("residuals_plot.png")
```

Appendix B: Raw R Code