The Effects of Distance on College Admission Emails

Michael Gillis
Emmanuel College
MATH 2113 Statistics with R
gillism@emmanuel.edu

December 16, 2022

1 Introduction

Each year, millions of prospective college students receive comparatively as many emails from colleges across the United States.[1] How do colleges determine who to send these promotions to? What factors lead to receiving an email from a college?

Since the beginning of the Information Age, colleges and universities across America have looked to make use of a variety of different digital methods to find and reach potential students.[2] While colleges still use the traditional methods of college fairs, snail mail, and outreach programs, these institutions have turned to opt-in email lists built from standardized tests, online interactions, and online tracking methods on certain websites.[2]

In one instance, a college used HTTP cookies, a digital tracking device placed on a computer that visits a website, in order to find students to whom they could target emails.[3] The University of Wisconsin-Stout used these cookies and an external online tracker that was able to recognize the computer of the student that was viewing their site. If they had previously applied or given any information about themselves to that college, the student would be targeted to receive increased emails from UW-Stout. This tracker would send the school's assistant director of admissions a report of any information that the site could find, including name, email, and an imprecise geographical location. Colleges are building massive databases of student information to refine their admissions processes, and a byproduct of this practice is the massive amounts of electronic mail sent to high school students.[4]

This paper looks for an association between different characteristics of a college and the number of emails that were sent in a sample. Specifically, one characteristic of interest was the distance of the college from the student's hometown. This led to the central research question: Is the distance of a college from a student's hometown associated with the number of emails the student receives from that college? It makes sense that a college would be more interested in sending promotional materials such as emails to students who live closer to their campus. The hypothesis that states, as distance increases, email count will decrease, follows from this theory.

The paper is structured as follows: Section 2 describes the methods used during the research, Section 3 outlines the results of the analyses that were done, Section 4 discusses the findings and possible interpretations of the results, and Section 5 concludes the paper with references.

2 Methods

2.1 Data Source

This data was obtained by tracking emails sent to the inbox of a high school student in the graduating class of 2021. This data was collected from September 5th, 2019, to September 5th, 2021. Those two years marked the entirety of the student's Junior and Senior years of high school, as well as the summer after their graduation.

Each email was flagged, then recorded with the college name and the date sent. Later, other variables that corresponded to characteristics of the college that sent the email were recorded from a database on the website of the US News and World Report.[5]

The recipient email was a personal Gmail account. The student had selected “Yes” on a questionnaire from the College Board which asks students whether or not they would like to opt-in to their “Student Search Service”, which provides information about the associated student to an extensive list of colleges.[6] The email used in this research was included in the student’s profile for the Search Service.

2.2 Variables & Statistics

The variables collected from the email inbox and through the US News college search database were college name, email count per college, distance, average grade point average (GPA), acceptance rate, enrollment size, and endowment. While dates were collected, they played no role in the analyses of the data.

College Name: The name of the institution that sent the email is used as the identifier variable, meant only to differentiate between entries in the data.

Count per College: The number of emails sent by each college was tracked as a discrete quantitative variable taken from the inbox, and is used as the response, or dependent variable throughout this research.

Distance: The distance from the student’s hometown was recorded in miles, taken from the U.S. News database. This is the main predictor, or independent variable throughout this research.

Average GPA: This continuous variable is a measure of the average GPA (4.0 Scale) held by the students at that school, based on the U.S. News database. This is not used in the multivariable model, for reasons explained later.

Acceptance Rate: This variable is a proportion of the number of students accepted divided by the number of students who applied, rounded to two decimal points from the U.S. News database. This is used as an alternate predictor variable in the multivariable model.

Enrollment Size: The number of undergraduate students enrolled was recorded as a discrete variable from the U.S. News Database. This is used as an alternate predictor variable in the multivariable model.

Endowment: The amount, in millions of dollars, that a college has in their endowment, was tracked as a continuous variable rounded to two decimal places from the U.S. News database. This is used as an alternate predictor variable in the multivariable model.

2.3 Statistical Analysis

All variables in this study besides the identifier were quantitative, and they were summarized by their medians and 1st and 3rd quartiles, which make up their interquartile range. A simple linear regression model is used to analyze the relationship between distance and count, and a multivariable linear regression model is used to analyze the relationship between the same variables, while controlling for variables that are statistically significant at the level 0.10. Significance is denoted as $p \leq 0.05$. One-tailed tests are run (p-values divided by 2), to test the hypothesis of a negative correlation. RStudio "Spotted Wakerobin" Release (7872775e, 2022-07-22) for macOS was used for data analysis.

3 Results

The results section consists of (3.1) a table of the summary statistics for each variable, including histograms and boxplots for the two main variables (count and distance), (3.2) the results of the simple linear regression model of distance vs. count, and (3.3) the results of the multivariable linear regression model, which controls for the alternate predictor variables.

In the study, 153 different colleges sent a total of 7294 emails to the inbox over the two years of tracking for this research. One of these colleges was outside of the United States and is excluded from the analysis.

3.1 Summary Statistics

Variable (n = 152)	**Median (Q1, Q3)**
Count per College	42 (18, 65)
Distance	312 (137.5, 921.5)
Average GPA (4.0)	3.65 (3.448, 3.815)
Acceptance Rate (%)	66.9 (36.5, 77.9)
Enrollment Size	5578 (2700, 10840)
Endowment (mil. $)	474.5 (197.5, 1390.0)

Table 1: The median and interquartile range (Q1, Q3) for each quantitative variable.

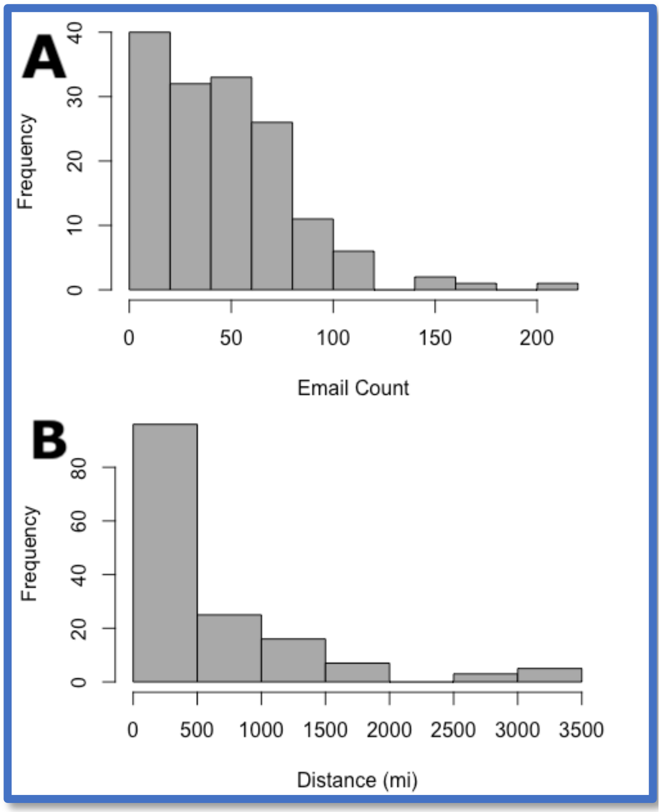

Figure 1: Histograms showing the distribution of the two main variables, **(A)** count (outcome), and **(B)** distance (main predictor).

Count per college has a median of 42 emails with an interquartile range of 18 to 65. Distance has a median of 312 miles with an IQR of 137.5 to 921.5 (Table 1). The distributions for distance and count were heavily skewed (Figure 1), so the Student's *t*-distribution is used in the analysis with $n - 2 = 150$ degrees of freedom.

3.2 Simple Linear Regression

Simple Linear Regression Outcome Variable: **Count**			
Predictor Var.	ß	*p*-value	R²
Distance	−0.00774	0.0208	0.0273832
Avg. GPA	−9.833	0.12904	0.0085187
Acceptance Rate	20.757	0.02815	0.0240742
Enrollment Size	−0.000406	0.0367	0.0212101
Endowment	−0.0010950	0.00468	0.0450456

Table 2: A table of values for each predictor variable using simple linear regression, listing both the coefficient value (ß) and the *p*-value.

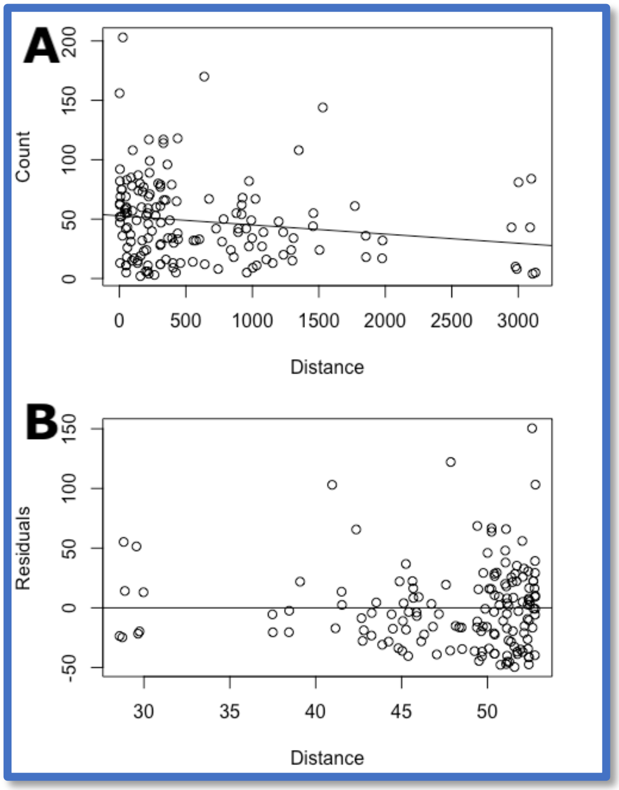

Figure 2: **(A)** Scatterplot of distance vs. count, with a least squares regression line. **(B)** Plot of residuals vs. predicted values for distance.

The simple linear regression model found a significant association, with $p < 0.05$, between distance and count when not controlling for other predictors. The R^2 value for this association is roughly 0.0274. The slope of the least squares regression line (ß) is approximately -0.0077 (Table 2).

The scatterplot of the association (Figure 2A) shows that the relationship is linear. The residual plot (Figure 2B) shows that the residuals are evenly distributed above and below the horizon line, with no clear pattern in the data points. Therefore, the assumptions for running a linear model are met.

Average GPA has an insignificant association with count, where $p \geq 0.10$ (Table 2), so it will be excluded from the multivariable analysis.

3.3 Multivariable Linear Regression

Multivariable Linear Regression Outcome Variable: **Count**			
Predictor Var.	ß	*p*-value	R^2
Distance	-0.0063433	0.0478	0.05882
Acceptance Rate	8.5450281	0.25715	
Enrollment Size	-0.0004038	0.03565	
Endowment	-0.0007970	0.0555	

Table 3: A table of values for each predictor variable using multivariable linear regression (controlling for all other significant predictors), listing both the coefficient value (ß) and the *p*-value.

The multivariable linear regression model found a significant association, with $p < 0.5$, between distance and count when controlling for other predictors. The R^2 value for this association is roughly 0.0588. The slope of the least squares regression line (ß) is approximately -0.0063. There is sufficient evidence to reject the null hypothesis that there is no significant correlation between distance and count.

The model also found that, when controlling for other predictors, enrollment size is significantly associated with count, with $p < 0.5$. All other predictor variables had p-values above the level of significance in the multivariable model.

4 Discussion

4.1 Interpretation of Models

Both the simple and multivariable linear regression models found a significant association between distance and count. Because the test performed was one-tailed, the significance of this association indicates that there is a negative correlation between the two variables.

Indeed, the regression coefficient (ß) is < 0, so the slope of the least squares regression line (LSRL) is negative. As the distance of a college from the student's hometown increases, the predicted email count for that college decreases.

The R^2 value for this correlation is around 0.0274, which means that about 2.74% of the variability in the email count is explained by the simple model. This is quite low, so to account for more variability in email count per college, the multivariable model was run with significant alternative predictor variables.

Once again, in the multivariable model, the regression coefficient (ß) is < 0, so the slope of the LSRL is negative, which supports the same conclusion from the simple regression model. This time, the R^2 value is around 0.0588, which means that about 5.88% of the variation in email count can be explained by the linear model.

Enrollment size was the only other variable that had a significant relationship with count. It also had a ß < 0, supporting the conclusion of a negative correlation. This means that, as the enrollment size of a college increases, the predicted value of the count decreases.

4.2 Drawbacks and Limitations

The research could have been improved with multiple students, selected randomly as to get a representative sample of geographic location, academic performance, and demographics. The student whose inbox was tracked for this research lives in a city with a high number of colleges per capita, so acquiring a representative sample of students, especially for a study involving distance and location, could improve this analysis.

Dates were tracked, but not used in the analysis. There was no variable that could be gleaned from the range of dates that could easily be attached to each college, so the variable was left out. If some method of quantifying this list of dates into a singular variable was found, then it could be added to the model and accounted for.

The data was tracked manually, so errors and missed entries were bound to have happened. A program that scrapes the inbox for specific emails could be written and could automate that process.

4.3 Application to Research Question

The outputs of the models suggest the existence of a significant correlation between distance and count. There is evidence to suggest that colleges are looking to market towards students whose hometowns are closer to their campus. The correlation found in the models is not extremely strong, so more evidence would need to be presented to come to a stronger conclusion about this correlation.

Future research could possibly involve studies with college admissions counsellors and marketers, who design these outreach processes from scratch, with limited resources and an increasing number of ways to reach students.[1,2,4,7]

There is also evidence to suggest a possible correlation between enrollment size and count. Colleges with larger enrollment sizes are succeeding with attracting students, especially those with larger brand names.[8] This could possibly be one of the underlying factors behind this correlation, but much more research featuring this idea would need to be done about enrollment size and marketing strategies for colleges to come to any type of conclusion.

5 References

[1] Abdul-Alim, Jamaal. “Flipping the Admission Process.” *The Journal of College Admission* 256 (2022). https://www.nacacnet.org/resources/newsroom/journal-of-college-admission/.

[2] Blumenthal, Tricia. “Why Do I Get So MUCH College Mail?” Vanderbilt University. Vanderbilt University, September 19, 2017. https://admissions.vanderbilt.edu/vandybloggers/2017/09/why-do-i-get-so-much-college-mail/.

[3] Fiebrandt, Stephan. “What Are Cookies? What Are the Differences between Them (Session vs. Persistent)?” Cisco. Cisco, July 17, 2018. https://www.cisco.com/c/en/us/support/docs/security/web-security-appliance/117925-technote-csc-00.html.

[4] MacMillan, Douglas, and Nick Anderson. “Student Tracking, Secret Scores: How College Admissions Offices Rank Prospects Before They Apply.” The Washington Post. WP Company, October 14, 2019. https://www.washingtonpost.com/business/2019/10/14/colleges-quietly-rank-prospective-students-based-their-personal-data/.

[5] U.S. News & World Report. “College Search: Find the Best Colleges & Universities.” U.S. News. Accessed December 12, 2022. https://www.usnews.com/best-colleges/college-search.

[6] “Student Search Service: Connect with Colleges and Scholarships.” College Board Big Future. Accessed December 13, 2022. https://bigfuture.collegeboard.org/student-search-service.

[7] *The 2019 State of Higher Ed Marketing*. Alexandria, Virginia: Simpson Scarborough, 2019. https://f.hubspotusercontent30.net/hubfs/4254080/The%20State%20of%20Higher%20Ed%20Marketing.pdf.

[8] Busta, Hallie. “Report: Size and Resources Will Determine Colleges' Success in the Next 10 Years.” Higher Ed Dive, July 9, 2019. https://www.highereddive.com/news/report-size-and-resources-will-dictate-colleges-success-in-the-next-10-ye/558353/.