

# Stats 504 Assignment 6: Graduation Rates

## Introduction

Admission grade cutoffs are used to determine if a student is admitted into a university or not. These cutoffs are important because they greatly impact both the university and the student. If the cutoffs are too low, students who will potentially struggle with the coursework and fail to graduate are admitted and this is unfavorable for both the student (who pays for tuition, and does not get a degree) and the university (whose graduation rates go down). The aim of this report is to investigate and shed light on the effect of admission grade cutoffs on the chances of a student graduating, given other details about the student. It is found that admission grade is indeed an influential factor (and specifically has a causal relationship) with the graduation rate, and more concrete findings are presented in the results section.

## Methods

The goal of the analysis is to help depict and understand the relationship between the admission grade and chances of graduation of a student. Other information about the student such as age, gender, parents' grades and occupations, previous qualification and corresponding grades are also used in determining this outcome.

Since this analysis involves identifying whether changes in admission grade cutoffs affect the chances of a student graduating, causal inference methodologies were adapted as this is more informative as compared to linear regression modeling for example, which only provides information about associations which are not necessarily the same as a causal relationship. To put it another way, the client wants to determine not just how a change in admission grade changes the chances of graduation, but also if that change is directly accounted for by the change in admission grade (causally). This is precisely what causal inference methods provide.

The primary idea in causal inference is to capture the effect of “confounding variables” so that the true causal parameter estimate can be computed. A confounding variable is one which affects both the treatment and outcome variables (which in this analysis are admission grade and graduation). If confounding variables are not controlled for, the association between the treatment and outcome will include the effect of the confounding variables. Thus, including the confounding variables in the model will capture their effects thereby isolating the effect of the treatment which is the true causal estimate. Another aspect of causal inference is the usage of “propensity scores” or weights for each observation. These weights are used to reduce selection bias and make the treatment and control group comparable to each other. In this analysis, the treatment variable is numerical (admission grade) so it is possible to split it into two categories based on the median value and use those as two treatment groups (this is presented in the baseline table of the results section). However, this results in loss of information so the decision was made to treat the admission grade as a continuous treatment variable, and use an adapted method to

calculate propensity scores for weighting each observation of the data, which is later used in the weighted linear model to compute the causal effect of admission grade on graduation status.

More specific to this analysis, thought was put into selecting confounding variables from the dataset and controlling for them in the analysis so that the true causal effect of admission grade on graduation is isolated. This variable selection is subjective, and based on intuition rather than quantitative tests, although t-tests and chi-square tests are used as ways of validating the intuitive reasoning for including or excluding a variable. The reasoning for dropping a few important variables that may look seemingly important to the client, are presented in the following paragraphs.

Application Mode and Application Order indicate how the student applied to the university, and since these occur after they take the admissions test (and assigned an admission grade), these variables are “downstream” variables, meaning they cannot affect admission grade. For this reason, they cannot be confounders, and hence were excluded from the analysis. Similar reasoning follows for the Course the student applied to study. It is likely that they decide the course based on their scores and a self-evaluation of how likely they are to be admitted to the program for a given course. However, their course selection cannot affect their admission grade, and thus was also excluded from the model. Tuition and Scholarships are also similar in that they are applicable only once the student is admitted to the university (and thus already has an admission grade). Also, all the information about the curricular units the student is enrolled in is excluded for the same reason that this is after they start school.

The data dictionary does not provide any information about the Displaced variable, and without more information it seemed reasonable to exclude the variable from the analysis to avoid any unwanted effects from including it.

Finally, the nationality variable was excluded and perhaps this is the most subjective assumption of all, but it seemed fairly reasonable to assume that a student would perform the same on an admissions test and in college regardless of their nationality. GDP, Inflation Rate, Unemployment Rate, and Age were excluded based on a t-test which showed that they had no significant effect on the admission grade, and it seemed reasonable to assume that these factors do not necessarily affect a student’s performance on an admissions test.

The causal inference model is fairly complex and has its share of limitations. It assumes that there are no latent confounders i.e variables which are unobserved but ones which are not captured by the dataset. This is a reasonable assumption but it is likely that the dataset does not have all the variables that affect both admission grades and graduation rates. However, with enough variables the model should be sufficient to provide a good enough estimate of the coefficients. Another assumption is that the variables have a linear relationship in the data. This is also unlikely to be completely true, as we don’t expect these variables to change linearly, but this assumption can be relaxed and the model will still be interpretable if the variables are not entirely linearly dependent.

## Results

The client provided the dataset required for this analysis as a CSV file which was acquired from several disjoint databases related to student enrollment in undergraduate courses between 2008 to 2019. This data represented various student information along with the status of whether they graduated, dropped out, or were still enrolled at the end of the stipulated time to degree. Each row of the data represents a single student, and there are 4424 such records. This data was already preprocessed by the client, so there was no necessity for cleaning or handling missing values. For each student, there was a multitude of information most of which proved fairly unusable for this specific type of analysis. If the variable was not a confounder it was discarded from the analysis, as described in the methods section. The causal inference analysis was finally performed with each observation having 16 different features. The numerical features are further described in Table 1, while the categorical features are presented in the appendix due to there being a large number of categories for each variable.

Features	Admission Grade (Lower Half)	Admission Grade (Upper Half)	p-value
Age at Enrollment	20 (19, 25)	20 (18, 25)	0.8
GDP (% change)	0.32 (-1.70, 1.79)	0.32 (-1.70, 1.78)	0.094
Unemployment Rate (%)	11.10 (9.40, 12.70)	11.10 (9.40, 13.90)	0.3
Previous Qualification Grade	127 (120,133)	140 (132, 147)	<0.001 *
Inflation Rate (%)	1.40 (0.30, 2.60)	1.40 (0.30, 260)	0.5

\* Significant Feature

**Table 1:** Baseline table indicating summary statistics of data of numerical features

The regression model was fitted, and the results are presented in Table 2. More insights which may be useful to the client are presented in Table 3.

Features	Coefficient	p-value
Admission Grade: Dropout	-0.018	<0.001
Admission Grade: Enrolled	-0.017	<0.001

**Table 2:** Regression Coefficients. The two features are interaction terms between graduation status and admission grade (with graduate as reference group.)

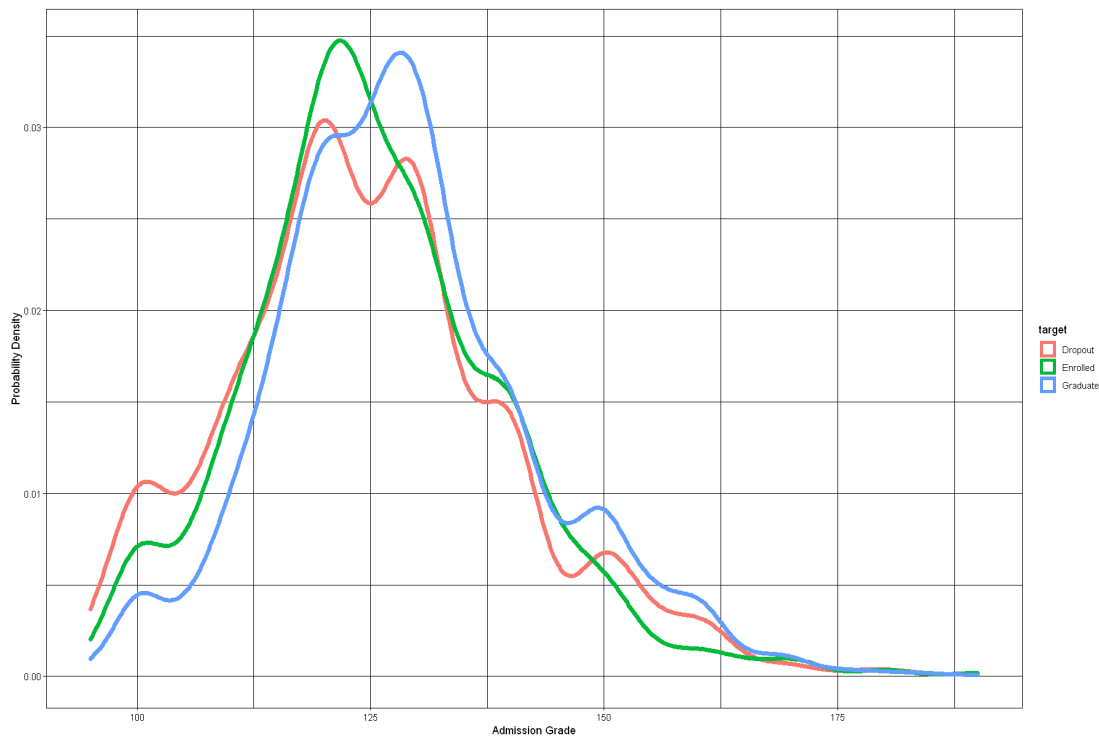
With a one point increase in admission grade, chance of dropout (Admission Grade: Dropout) goes down by 1.8% and chance of still being enrolled (failed to graduate in stipulated time) goes down by 1.7%. These values are in comparison with graduating, if all other variables are held constant.

In order to shed additional light for the client, the model can be used to compute the expected probability change of the graduation status given different admission grade cutoffs for which the client is contemplating a change. The following table presents some realistic scenarios for the client to consider.

Change in Admission Grade	Decrease in Dropout (%)	Decrease in Enrolled (%)
+5	8.75	8.35
+10	17.5	16.7
+20	35	33.4

**Table 3:** Changes in graduation status with different levels of admission grades.

Additionally, Figure 1 below provides information on how admission grades are distributed amongst the different graduation status. For example, the admission grade of a majority of students who graduated is higher (blue peak) than that of those who are still enrolled and those who dropped out of university (green and red peaks). In reference to the client's question about where to set the admission grade cutoff, a reasonable level seems to be 125 points, relative probabilities of graduating are higher than dropping out or staying enrolled as is depicted in Figure 1. If the client desires a more lax cutoff, they may refer to the figure to make their decision.



**Figure 1:**  
Density Plot  
for Admission  
Grade across  
Graduation  
Status

## Conclusion

This report presents the results of a causal inference analysis performed on student information, with the objective of trying to understand the effect of admission grade on the graduation status of the student. Admission grade is deemed a significant variable by the model and impacts the graduation rate in a causal fashion. An increase in one point of admission grade, lowers the probability of dropping out or staying enrolled at the end of the stipulated time to graduation by 1.8% and 1.7% respectively, as compared to the probability of graduating. While these results are expected to be useful, it must be noted that the model comes with a fair share of limitations; It assumes that all confounding variables are included in the model, and that there are no latent (i.e unobserved) confounders. It also assumes linear relationships in the data, and the propensity score weights are accurate. While there may be limitations in the methods, it is expected that these results will still guide the client in tinkering with their admission grade cutoffs in order to enhance both their reputation with respect to graduation rates, as well as providing students a fair opportunity to earn a degree.