The Causal Effect of Student Absences on Mathematical Educational Performance Using Linear Regression and Inverse Propensity Weighting

We analyzed the effect of student absences on mathematical educational performance in Portugal from the UCI Machine Learning Repository to better understand how absence affects educational performance in general. Specifically, we were interested in how three or a greater number of absences affected final math scores. While the data only included two schools in a different country, this analysis would provide preliminary information on what to expect of the relationship between absences and performance for further research. Using linear regression, and inverse propensity weighting (IPW), we discovered that having greater than three absences decreased a student's math score by 1.13.

The data set provided a multitude of variables outside of absences and final math scores. We chose variables that might introduce a confounding relationship between absences and math score. Linear regression might pick up on the relationship from these confounding variables, which affect both variables of interest outside of those variables of interest, without distinguishing that confounding relationship as separate from the relationship between only absences and final math scores. Table 1 enumerates how different each variable is for students with less than three absences and students with greater than or equal to three absences. We gave measurements of the 2 groups for each variable as well as the p-values from a chi-square test that measures how likely the 2 groups are the same. We noticed a difference between the 2 groups for a number of variables, confirmed by the small p-values. The variables with the lowest p-values were age and how often the student went out with friends, meaning that these variables had the greatest disparities between the two absence groups. Since there were differences between the absence groups, we needed to take these differences into account before analyzing the causal relationship between absences and math score.

Linear regression was the straightforward model for analyzing the relationship between absences and math scores, but first we need to eliminate any other variables that might affect these two variables of interest to justify a causal relationship. This causal relationship would describe how absences affect math scores given that every other factor about the student was exactly the same. We used random forest boosting to calculate inverse propensity weights as our method of eliminating the effect of the other variables shown in Table 1. So we calculated inverse propensity weights by first categorizing each student observation into groups described by the baseline characteristics of Table 1 and then balancing the impact of that observation based on how many student observations were in each group. The end result of IPW were decreases in differences between the students with less than three absences and students with

greater than are equal to three absences, in effect nullifying the influence of all the other variables except for the variables of interest.

After we calculated the proper weights and used them to nullify other variable effects, we used a simple linear regression model of absences on math score given the IPW analysis to determine the causal effect of absences on math scores. Table 2 includes the results of that model. Assuming every other factor for the student was the same, the effect of having three or more than absences decreased the student's math score by 1.13 compared if that student had less than three absences. In general, more absences led to decreased math scores.

In conclusion, we wanted to analyze the effect of student absences on mathematical educational performance by modeling the causal relationship between greater than three absences on student final math scores. We assumed that the other variables available in the data set included all variables that could possibly influence both these two variables of interest and that all the variables included did have a confounding influence on these two variables of interest. We also assumed that these students were representative of all the students in Portugal as well as all students in general. If all these assumptions were to hold true, we could conclude that greater than 3 student absences decreases mathematical educational performance in general, but it's unlikely that all of these assumptions were true. Other confounding variables could be a student's previous academic aptitude and mental health. Since Portugal and the United States, nevermind the specific population of the data set and the students of Ann Arbor, are different in demographics as well as economics and politics, it's difficult to argue that the two are similar enough that this data set was representative of students in America. The exact numerical effect of absences on educational performance may be different for students in America from that of students in the data set, but, since this was a preliminary analysis, the relationship found in the analysis can be a piece of the overall analysis.

Table 1: Baseline Characteristics for each Covariate (Percentage % for Binary and Mean(Standard Deviation) for Continuous Variables)

Variable	Total (n=357)	Absences less than 3 (n1=145)	Absences greater or equal to 3 (n2=212)	p-value
School (MS)	88.2%	87.6%	88.7%	0.883
Sex (F)	51.8%	51.0%	52.4%	0.890
Age	16.7(1.27)	16.3(1.14)	16.9(1.29)	0.003
Rural	21.8%	19.3%	23.6%	0.407
Famsize (GT3)	70.0%	71.7%	68.9%	0.645
Pstatus (A)	10.9%	7.6%	13.2%	0.134
Schoolsup (No)	86.0%	84.8%	86.8%	0.711
Famsup (No)	38.7%	37.9%	39.2%	0.903
Activities (No)	49.6%	52.4%	47.6%	0.437
Nursery (No)	19.9%	20.0%	19.8%	1.000
Higher (No)	3.9%	4.1%	3.8%	1.000
Internet (Y)	16.2%	17.9%	15.1%	0.570
Romantic (Y)	68.6%	74.5%	64.6%	0.063
Famrel	4.0(0.89)	4.1(0.76)	3.9(0.96)	0.225
Free Time	3.2(1.01)	3.2(0.97)	3.2(1.04)	0.863
Go Out	3.1(1.09)	2.8(1.05)	3.3(1.08)	0.004

Dalc	1.5(0.92)	1.3(0.61)	1.6(1.06)	0.007
Walc	2.3(1.29)	2.0(1.18)	2.5(1.33)	0.006
Health	3.5(1.40)	3.7(1.44)	3.4(1.37)	0.011
Mother Higher Educ (No)	65.0%	64.8%	65.1%	1.000
Father Higher Educ (No)	75.4%	73.8%	76.4%	0.660
Close to School(No)	33.9%	34.5%	33.5%	0.936
Math Score	11.5(3.23)	12.3(3.04)	11.0(3.25)	0.025

Table 2: Coefficients from Linear Model of Absence on Final Math Score

	Estimate	**p-value**
Intercept	12.2	< 2e-16
3 or greater absences	-1.13	0.0008