Effect of Absences on Students' Math Grades

This analysis aims to understand the effect that three or more absences (treatment group) in a school year has on final math grades using data from two high schools in Portugal. We measure the specified impact by estimating the average treatment effect (ATE), or the difference in mean final math score for students in the treatment group and the mean score for those in the control group (less than three absences). We use propensity scores calculated through a boosted logistic regression model in order to balance out the distributions of covariates in the two absence groups. We also identify demographic variables and school-related attributes that affect the number of absences by building a quasi Poisson model. For both tasks, we consider the following predictors: the student's school, travel time to school, sex, age, health status, whether they have an interest in purusuing higher education, and whether they have extra educational support at home or in school. We found that having three or more absences cause students' math scores to be 1.785 points higher, and that the older a student's age, having extra educational support at school, and attending the Gabriel Pereira School are associated with increases in the number of absences

The table below shows that for the different covariates we are considering, their distributions are slightly different when comparing the treatment and control groups. For example, in the treatment group, 96% of students are interested in higher education versus 93% in the control group. Since our focus is to identify the causal impact of three or more absences on math scores, these different proportions can lead to a misleading average treatment effect. To accurately measure the treatment effect, we can mimic a randomized controlled study by weighting each observation by an amount such that for both the treatment and control group, we end up with similarly distributed numerical variables, or categorical variables with similar subgroup sizes for each possible category. We note that performing the Wilcoxon rank sum test and the chi-square test for numerical and categorical variables respectively show that most variables are independent of the treatment/control group, however we further improve the distributions with weighting. The weights correspond to 1 divided by the probability that that student had three or more absences given the values of his/her covariates.

We use the adaboost algorithm with logistic regression, which predicts the probability of a student being in the treatment group gives his/her covariate values. Age and whether or not the student had educational support from their family contributed the most to the treatment and control groups being the most unbalanced. We note that there is an overlap in the probability of being and not being in the treatment group given the covariates, an assumption that must be satisfied in order for us to use the weights to determine an unbiased ATE. To assess the covariates' balance after weighting, we calculate the absolute standardized difference (ASD) by subtracting the means of each covariate in the control and treatment group, and dividing by the standard deviation of that covariate. Our weighting was relatively successful in balancing the groups as all of our covariates ASD's are less than 0.07.

Using the weights for each observation, we calculate the average treatment effect by finding the average math final scores for students in the control and treatment groups separately: for each group, we sum the products of each student's final math score and weight value, and divide by sum of the weights of everyone in that group. On average, those with three or more absences will have a final math score that is 1.785 points higher compared to those with less than three absences. This counterintuitive result may be unique to the relatively small group of students in the dataset, and further analysis using a larger dataset can be done to verify if the direction of our estimated effect is common among students. One explanation for our result may be that for students in the two schools that we considered, they may value studying on their own at home more than learning at school, so they are less concerned with missing school. The p-value for the ATE, 0.0026, gives evidence supporting our obtained ATE as opposed to a value of 0. The result found here is accurate assuming that there are no other confounders that could impact both the final math score and whether a student has three or more/less than three absences.

To determine the effect of variables on the number of absences, we build a quasi Poisson regression model with the number of absences as the outcome variable and include the same predictors as above. We found statistical evidence of an association between the number of absences and age, sex, school, living in an urban versus rural area, and having extra educational support at school. Given the same values for all other variables, the expected number of absences for someone who is a year older increases by 28%. Compared to females, males have a 14% decrease in the expected number of absences, and those living in urban areas had a 15% decrease in absences compared to those living in rural areas. Students attending Mousinho da Silveira School have 57% less absences than those at Gabriel Pereira School, and students with extra support at school had 24% more absences compared to those who did not. Since our outcome variable includes all absences counted at the end of a year, it is plausible to assume that the covariates have a causal effect on the number of absences, instead of a causal relationship in the opposite direction.

Our estimated ATE of three or more absences in increasing final math score by 1.8 points seems accurate since we considered a variety of variables that likely affect both the number of absences or math final score directly. However, we may not have captured the true impact of the treatment if there exist any other confounders that affect being in the treatment/control group or math score. The same assumption must also apply in order for the identified effects of variables impacting the number of absences to be reliable.

Table of Covariate Distributions

Variables	**Total**	**Treatment (three or more absences)**	**Control (less than three absences)**	**P-value**
Age (years)	17 (1.276)	16.896 (1.295)	16.464 (1.217)	0.0010
Sex (m or f)	47.342% (187) 52.658% (208)	47.642% (101) 52.359% (111)	46.995% (86) 53.005% (97)	0.9782
Health (1 (very bad) to 5 (very good))	11.899% (47) 11.392% (45) 23.038% (91) 16.709% (66) 36.962% (146)	11.792% (25) 14.151% (30) 23.585% (50) 18.868% (40) 31.6035% (67)	12.022% (22) 8.197% (15) 22.404% (41) 14.208% (26) 43.169% (79)	0.0934
Interested in Higher Education (y or n)	94.937% (375) 5.063% (20)	96.226% (204) 3.774% (8)	93.443% (171) 6.557% (12)	0.3038
Family Educational Support (y or n)	61.266% (242) 38.734% (153)	60.849% (129) 39.151% (83)	61.749% (113) 38.251% (70)	0.9367
Extra Educational Support at School (y or n)	12.911% (51) 87.089% (344)	13.208% (28) 86.792% (184)	12.568% (23) 87.432% (160)	0.9693
School (Gabriel Pereira or Mousinho da Silveira)	88.354% (349) 11.646% (46)	88.679% (188) 11.321% (24)	87.978% (161) 12.022% (22)	0.9527
Travel time (1 (<15 min), 2 (15 to 30 min), 3 (30 min to 1 hr), 4 (>1 hour))	65.063% (257) 27.089% (107) 5.823% (23) 2.025% (8)	66.509% (141) 25.472% (54) 6.132% (13) 1.887% (4)	63.388% (116) 28.962% (53) 5.464% (10) 2.186% (4)	0.8715

The table above shows the distribution of covariates used in this analysis. For the numerical variable 'age', we include the mean and standard deviation, and the p-value for the Wilcoxon rank sum test to determine whether the two groups come from the same population. For the categorical variables, we include the percentages and counts, and the p-value comes from the chi-square test to determine whether the variable is independent of the control/treatment group.