Isabelle Yang

Effect of Alcohol Consumption and Family Support on Student Absences

Introduction

The topic I chose is the effects of alcohol consumption and family educational support on student absences. Specifically, I wanted to examine the effects that weekday alcohol consumption, weekend alcohol consumption, and family educational support have on student absences.

Data

The data comes from kaggle.com. Kaggle is an online community of data scientists and machine learning engineers that offers a diverse library of public datasets for all users. The specific dataset I used was obtained in a survey of 395 students in Portugal who are enrolled in a math course in secondary school. Furthermore, as a note, there may be concerns about selection in terms of who shows up in my sample as some students who drink more alcohol or receive less family educational support may choose to not participate in the study.

In terms of cleaning, although there are many columns containing information related to social, gender, and study information about students, I limited my focus to weekday alcohol consumption, weekend alcohol consumption, and family educational support. There were no NA values, so I did not remove any data. Furthermore, the family educational support variable, famsup, originally outputted one of two responses: yes or no. In order to clean it for my regression, I transformed it into a binary variable that outputs 0 if the response is "no" and outputs 1 if the response is "yes." In addition, it is important to note the construction of the weekday alcohol consumption (Dalc) and weekend alcohol consumption (Walc) variables. They are both numeric variables that range from 1 to 5, with 1 being very low consumption and 5 being high consumption.

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Absences	395	5.709	8.003	0	0	8	75
Weekday Alcohol Consumption	395	1.481	0.8907	1	1	2	5
Weekend Alcohol Consumption	395	2.291	1.288	1	1	3	5
Family Educational Support	395	0.6127	0.4878	0	0	1	1

Table 1 shows the summary statistics of the variables in the dataset. The absences variable has a mean of around 5.7 absences with a positively skewed distribution and outliers that range up to 75 absences. Weekday Alcohol Consumption, Weekend Alcohol Consumption, and Family Educational Support are all self-reported by students who participated in the study. Both the Weekday and Weekend Alcohol Consumption variables range from a minimum of 1 to a maximum of 5 with Weekend Alcohol Consumption having a greater mean (2.291) than Weekday Alcohol Consumption (1.481). The dummy variable, Family Educational Support, has a mean of 0.6127 which shows that there are slightly more students who receive family educational support in this study.

Regression

To investigate the determinants of student absences, I run the following regression:

$$\text{absences} = \beta_0 + \beta_1 \text{Dalc} + \beta_2 \text{Walc} + \beta_3 \text{famsup} + u$$

where Dalc refers to weekday alcohol consumption and Walc refers to weekend alcohol consumption. Both variables range from 1 (very low consumption) to 5 (high consumption). famsup refers to family educational support, and it is a dummy variable for whether the student receives family educational support (yes or no).

The goal of this regression is to investigate the determinants of student absences. The two main questions I wanted to explore are (1) is weekday or weekend alcohol consumption a more significant determinant of student absences? And (2) does having family educational support have a significant effect on student absences? By including the Dalc, Walc, and famsup variables in my regression, I can capture both the individual behaviors (weekday and weekend alcohol

consumption) and external factors (family educational support) that contribute to student absenteeism.

Results

Table 2: Regression Results—Student Absences

	Dependent variable:
	Absences
Weekday Alcohol Consumption	0.353
	(0.429)
Weekend Alcohol Consumption	0.708[*]
	(0.407)
Family Educational Support	0.582
	(0.871)
Constant	3.208[***]
	(1.191)
Observations	395
R^2	0.021
Note:	[*]p<0.1; [**]p<0.05; [***]p<0.01

Table 2 shows the results from the regression. It appears that although higher weekday alcohol consumption and having family educational support are both correlated with more student absences, their estimated coefficients are statistically insignificant. On the other hand, higher weekend alcohol consumption is correlated with more student absences, and its estimated coefficient is statistically significant at the 0.1 significance level. It is interesting to see that the estimated coefficients of weekend alcohol consumption and weekday alcohol consumption are different, with one being statistically significant and the other being statistically insignificant. These findings suggest that even though there is a relationship between alcohol consumption and student absence, timing (weekday or weekend) of alcohol consumption may also play an important role.

Another noteworthy observation is the R^2 value of 0.021. This means that approximately 2.1% of the variability in the student absences variable can be explained by the independent variables, weekday alcohol consumption, weekend alcohol consumption, and family educational support.

Because the R^2 value is extremely low, it suggests that the independent variables explain a small portion of the variability in student absences; therefore, there may be other unobserved factors that explain the remaining majority of the variability such as family income, peer influences, romantic relationships, study time, health status, parent cohabitation status, etc.

Because my regression only includes weekday alcohol consumption, weekend alcohol consumption, and family educational support, we may be worried about omitted variable bias. In other words, unobserved factors like the ones listed above are not factored into my analysis but may be correlated with both the independent and dependent variables in my regression. For example, when study time, which is negatively correlated with student absences (students who study more have less absences from school), is omitted from the regression, this could lead to an overestimation of the effects of weekend and weekday alcohol consumption and family educational support on student absences.

There may also be concerns related to endogeneity, which happens if the independent variable is related to the error term. In my regression, it is very possible that in addition to the independent variables influencing the dependent variable, the dependent variable, student absences, influences the independent variables. For example, students who skip class more often may be more likely to drink more alcohol. Alternatively, students who skip class more often may receive less family educational support. These instances of reverse causality may result in endogenous independent variables, meaning their values may be correlated with the error term, potentially resulting in biased estimates of their coefficients.

Conclusion

My regression aimed to explain some of the determinants of student absences. Interestingly, I found that the independent variables I chose for my regression explain an extremely small portion of the variability in student absences, with weekday alcohol consumption and family educational support having statistically insignificant estimated coefficients. I also found that the estimated coefficient of weekend alcohol consumption was greater and more statistically significant than that of weekday alcohol consumption, indicating that timing of alcohol consumption may play a role in determining student absences. Furthermore, there are many

unobserved factors that contribute to student absences, so further analysis is necessary to fully understand the main determinants of student absences.