Predicting Academic Performance of High School Students

I. Introduction

As of 2019, over 7 million American 12-to 20-year-olds reported past month alcohol use, over 4 million reported past month binge drinking, and nearly 1 million reported past month heavy alcohol use (2019 National Survey on Drug Use and Health). The use of alcohol at an early age and at such an important time in one's academic career can have bigger impacts than the students may expect. According to a study done in 2011, there is clear evidence that there is an impact on the grades of high-school students who consume alcohol as well as an increase in the amount of difficulties that they experience in the classroom (Balsa, A. I., Giuliano, L. M., & French, M. T. (2011)). In addition to analyses of the effects on current performance, studies have attempted to model the risk factors leading to alcohol abuse among adolescents. Bahr et. al. find that strong family ties, peer influence, and long-term educational commitments have significant, though small, effects on an adolescent's likelihood to drink heavily (Bahr, S. J., Marcos, A. C., & Maughan, S. L. (1995)). Despite abundant literature on the impact of alcohol consumption on academic performance, there is a lack of research on the reverse relationship: the ability to predict alcohol consumption using academic performance. Using data from a survey of high school students conducted by the University of Minho, the capacity to predict an adolescent's alcohol consumption combining many of the risk factors identified by Bahr. et al will be assessed. In essence, to what extent do the same demographic indicators used to predict alcohol consumption have a differential effect when incorporating school performance?

II. Data and Methods

To answer the research question, data originally collected in 2008 by Paulo Cortez, a researcher at the University of Minho in Portugal, was used for our analysis. Cortez collected this data in hopes of finding a correlation between alcohol use of students in high school and any changes in academic achievement as a result of it. The subjects of the study were 634 randomly selected Portuguese high school students from two different schools, sampling data on personal characteristics and grades in their math and Portuguese classes. There are several variables defined in this data set. These include each student's school, sex, age, address, family size, parents' marital status, parents' occupations, parents' education, home-to-school travel time, weekly study time, number of past class failures, educational support, extracurricular activities, internet access, relationships, free-time, health status, school absences, workday and weekday alcohol consumption, and grades from their first period, second period, and final grade for each class. Grade performance (*G1,G2,G3*) is used as a principal variable of interest alongside many of these controls to predict a student's level of weekend alcohol consumption (*Walc)*. One issue observed in the data is that some of the students were repeated if they were surveyed in both their math and Portuguese classes. Students were unable to be uniquely identified, so data was limited to one class (math) to ensure that each of the data points was unique and not repeated.

Given this study's goal of determining how student performance differentially predicts alcohol consumption when interacting with other demographic predictors, the data was examined for general trends in interaction. According to the graph in Appendix I, it is clear that sex has a relative effect on final grade and weekend alcohol consumption with males exhibiting relatively less consumption with higher grades as compared to females. This relationship will be explored further on in the analysis. It was also discovered that there were several individuals who had a final grade of approximately 0, which would affect the model, as these grades are significantly lower than the average grades of the students. To address this issue, we removed the data points for those students whose grades were 0 (below 5) for a total of 357 observations.

III. Analysis

This study models the effect of numerous demographic indicators and school performance on weekend alcohol consumption. The response variable in the model, *Walc*, is an ordered categorical variable estimating weekend alcohol consumption, increasing from 1 to 5. *Walc* was preferred over *Dalc*, daily alcohol consumption, since the vast majority of respondents reported low daily alcohol consumption. With an ordinal response variable, a proportional odds (PO) model will ultimately be used.

The respondents to the survey answered a variety of questions, ranging from family size to the number of past class failures or final grade performance. Many of these, such as past class failures and final grade performance, are likely correlated (correlation coefficient = -0.294) or insignificant in determining *Walc*, and will thus be eliminated from the model. A baseline linear regression model with all variables was then estimated, and using the Akaike information criterion (AIC) as a method for variable selection, the combination of predictors that would generate the minimum AIC was calculated (AIC = 1058.031). This model is shown below.

$$\begin{aligned} Walc_i = {} & \beta_0 + \beta_1 sex + \beta_2 address + \beta_3 famsize + \beta_4 studytime + \beta_5 paid + \beta_6 nursery \\ & + \beta_7 famrel + \beta_8 goout + \beta_9 health + \beta_{10} absences + \beta_{11} G3 + \varepsilon_i \end{aligned}$$

The model reflects where *sex* is binary for male, *address* is binary for urban or rural home setting, *famsize* is binary for a family larger than 3, *paid* is binary for additional tutoring being paid for, *studytime* is ordinal for time spent studying per week, *nursery* is binary for having attended nursery school, *famrel* is ordinal for quality of family relationships, *goout* is ordinal for frequency of a night out with friends, *health* is ordinal for health status, *absences* is numeric for number of school absences, and *G3* is an integer for final grade.

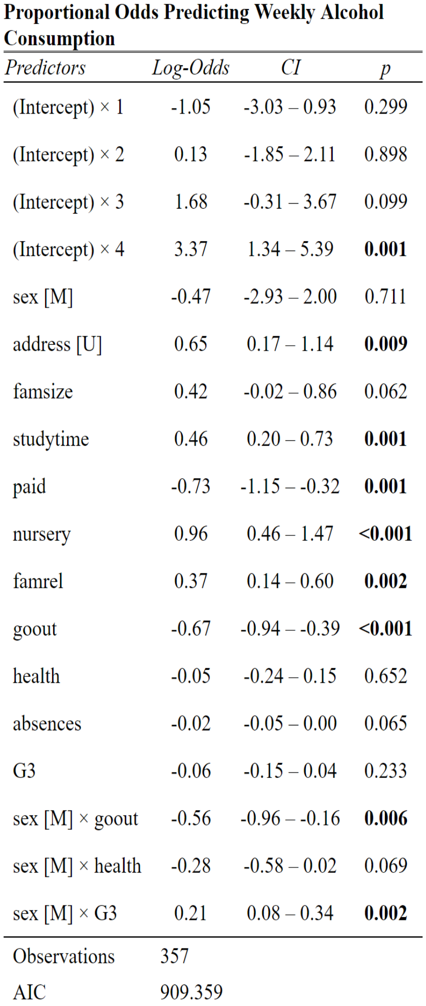
These predictors were used as benchmarks upon adopting a proportional odds model, where the same variables were used (AIC = 928.894). Our principal variable of interest, *G3*, was tested to be relevant to the model. Upon its omission, the AIC decreased only marginally (AIC = 928.814), so *G3* was left in the model due to it being integral to the research question. To account for nonparallel trends among the predictor variables, the Generalized Odds (GENORD) Model was also considered. However, due to additive complications from variables *studytime, paid, nursery,* and *address,* the nonparallel slopes were too interrelated to be calculated. As a check for robustness, both a PO and a GENORD model were calculated omitting these variables; however, the AIC value increased in both scenarios (956 and 976, respectively).

On the contrary, a goodness of fit test for the GENORD model for these less predictive models revealed that there is sufficient evidence to reject the PO model as adequate. However, the full PO model is used with the same predictors from the linear model since its AIC was lower. It would also be impractical to accommodate the full predictive power with the GENORD model due to the variables that result in additive complications.

To assess the differential effects of the indicator variables with grades, interactions of all variables with *G3* were included. Further given the relative effect of *sex* shown in the graphic in Section II, its interactions with the other predictor variables were also determined. The combination of interactions with the greatest predictive power was found to be *goout*sex, health*sex,* and *G3*sex.* All other interactions were dropped, and variables *sex* and *health* were kept in the model, despite being insignificant, to model the base level effect. The final, minimized AIC was calculated at 909.36. The following equation represents our final model.

$$\begin{aligned} & logit[P(Walc_i \leq j)] \\ & \quad = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \beta_1 sex + \beta_2 address + \beta_3 famsize + \beta_4 studytime + \beta_5 paid \\ & \quad + \beta_6 nursery + \beta_7 famrel + \beta_8 goout + \beta_9 health + \beta_{10} absences + \beta_{11} G3 + \beta_{12} sex \\ & \quad * goout + \beta_{13} sex * health + \beta_{13} sex * G3 + \varepsilon_i \end{aligned}$$

The model reflects where α_j is the associated log-odds of falling at or below a *Walc* level of *j*. The estimated coefficients and standard errors are shown below.



Proportional Odds Predicting Weekly Alcohol Consumption

Predictors	*Log-Odds*	*CI*	*p*
(Intercept) × 1	-1.05	-3.03 – 0.93	0.299
(Intercept) × 2	0.13	-1.85 – 2.11	0.898
(Intercept) × 3	1.68	-0.31 – 3.67	0.099
(Intercept) × 4	3.37	1.34 – 5.39	**0.001**
sex [M]	-0.47	-2.93 – 2.00	0.711
address [U]	0.65	0.17 – 1.14	**0.009**
famsize	0.42	-0.02 – 0.86	0.062
studytime	0.46	0.20 – 0.73	**0.001**
paid	-0.73	-1.15 – -0.32	**0.001**
nursery	0.96	0.46 – 1.47	**<0.001**
famrel	0.37	0.14 – 0.60	**0.002**
goout	-0.67	-0.94 – -0.39	**<0.001**
health	-0.05	-0.24 – 0.15	0.652
absences	-0.02	-0.05 – 0.00	0.065
G3	-0.06	-0.15 – 0.04	0.233
sex [M] × goout	-0.56	-0.96 – -0.16	**0.006**
sex [M] × health	-0.28	-0.58 – 0.02	0.069
sex [M] × G3	0.21	0.08 – 0.34	**0.002**
Observations	357		
AIC	909.359		

The results show that after controlling for a variety of relevant factors, the log odds and odds of being at or below a certain weekend alcohol consumption will decrease for males when compared to women but will increase significantly more for males who perform well (0.21x and 1.23x, respectively) than women. This aligns with the assumption that high-performing students can serve as a moderate indicator for drinking less, and gender serves as a strong indicator for alcohol consumption, especially when combined together. Other relevant predictors for drinking include the student's address being urban (*address*), time spent studying (*studytime)*, and healthy family relationship (*famrel*), all of which are positively correlated with probability of drinking less. Furthermore, going out has a noticeably large, negative effect among males.

The relative importance of grades by sex is demonstrated by the stacked probability graph in Appendix II. As grades increase, males exhibit a significant rise in probability of drinking less, whereas females show a more gradual decline, likely due to limits within the sample.

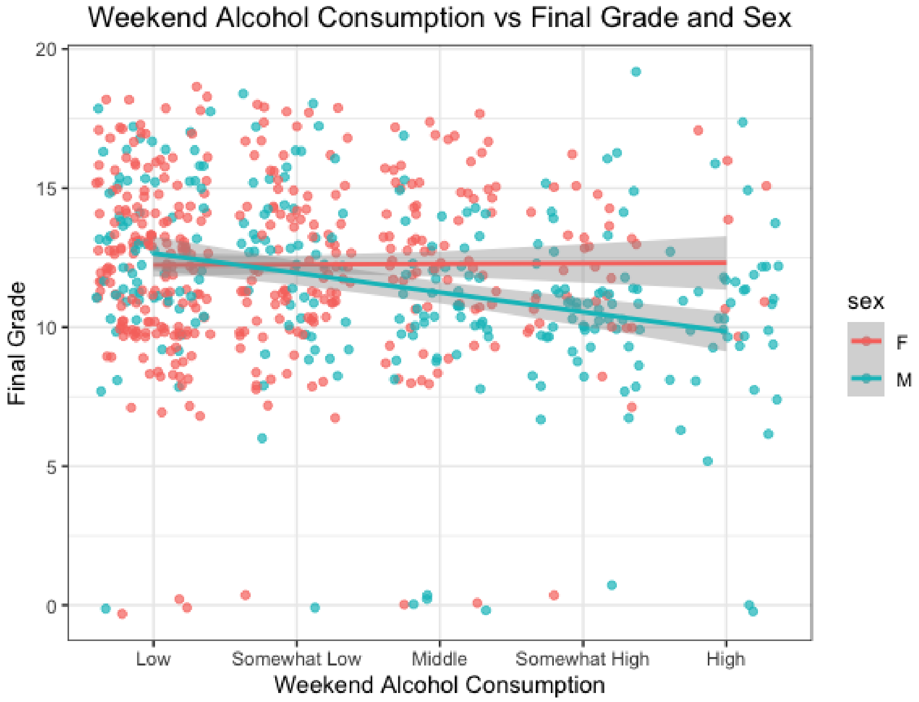
IV. Conclusion

This analysis answers the original question by providing evidence that alcohol consumption can be negatively correlated with academic performance and it tells that gender serves as a strong indicator for alcohol consumption for students in high school. The analysis determined that the weekend alcohol consumption for males is higher than that of females, but consumption decreases for males who perform better on exams compared to males with worse exam scores, whereas there was no such relationship for females.

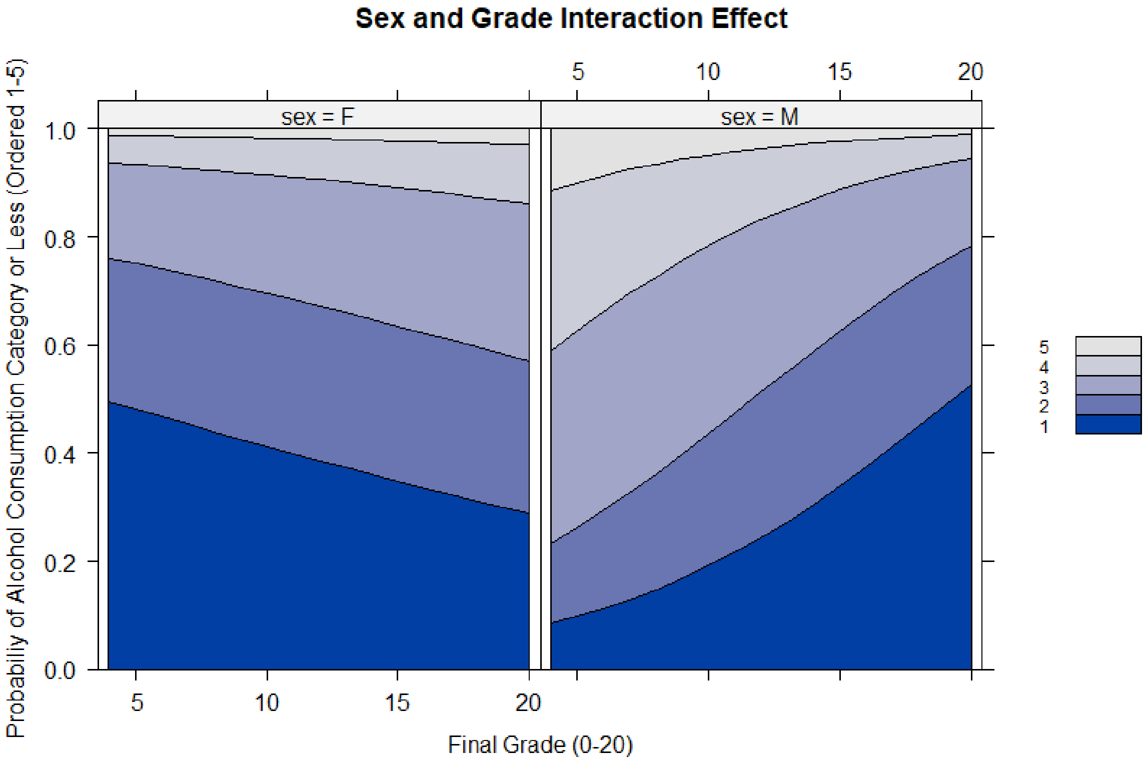
Furthermore, due to the response variable being an ordinal, categorical variable, the results were limited to a Proportional-Odds Model. Additional research and study would be needed to test for general normality assumptions, collinearity, and residual autocorrelation or to accommodate estimators to a GENORD model.. Such tests would validate the goodness of fit for the PO Model. Nevertheless, the model provides preliminary evidence that grades, in combination with other descriptors, serve as an effective indicator for the level of alcohol consumption. Taking this into consideration with previous research on alcohol consumption's effect on school performance, the implication could be the relative importance of both academics and social life for certain groups, compared to others. The research adds to the overwhelming literature on the importance of a healthy, prudent lifestyle, by identifying target audiences for awareness programs. In particular, both strong academics and healthy drinking habits should be strongly emphasized for males.

Appendix

Appendix I



Appendix II



Bibliography

Dataset:

UCI Machine Learning. (2016). Student Alcohol Consumption. Kaggle Inc. Student Alcohol Consumption | Kaggle

Sources:

Balsa, A. I., Giuliano, L. M., & French, M. T. (2011). The effects of alcohol use on academic achievement in high school.Economics of education review, 30 (1), 1. https://doi.org/10.1016/j.econedurev.2010.06.015

Bahr, S. J., Marcos, A. C., & Maughan, S. L. (1995). Family, educational and peer influences on the alcohol use of female and male adolescents. Journal of studies on alcohol, 56 (4), 457-469.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (2020). Results from the 2019 National Survey on Drug Use and Health: Detailed tables. Detailed Table 7.16A. https://www.samhsa.gov/data/report/2019-nsduh-detailed-tables