Alcohol Consumption in Schools

Hypothesis Heroes - Drew Davison, Lisa Zhang, Ellie Culman, Austin Chang

2023-11-14

Introduction and data

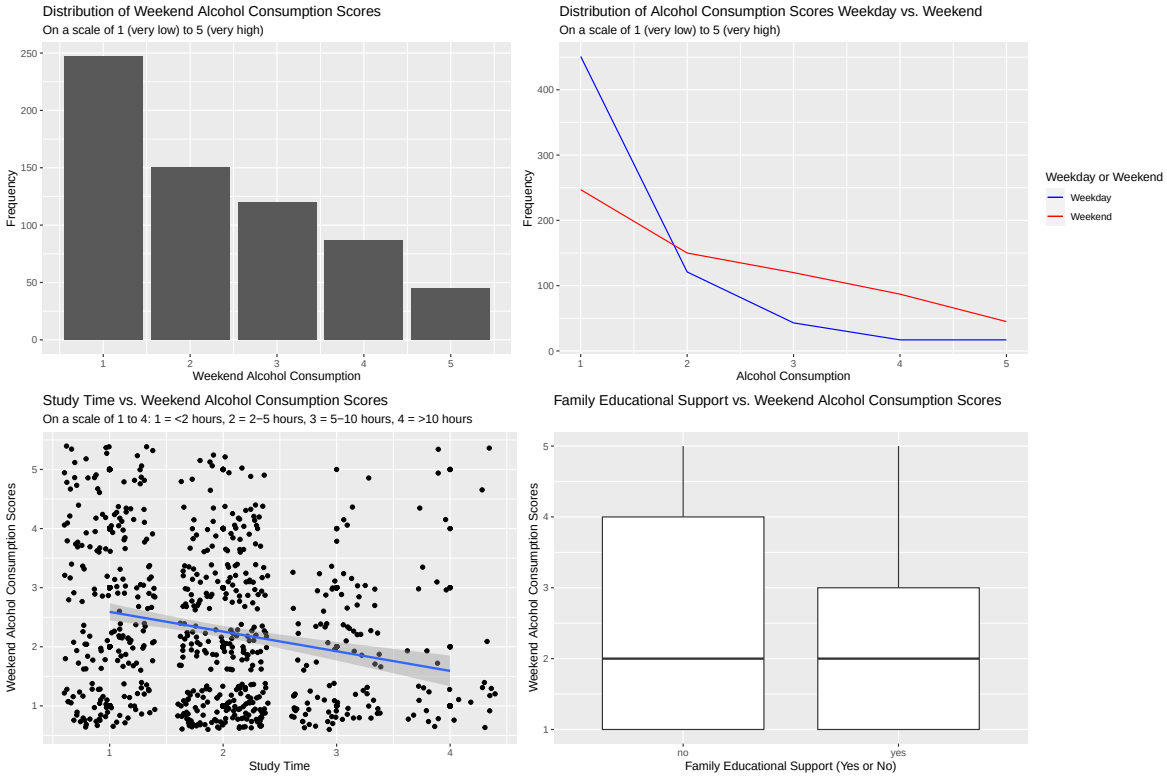
All around the world, underage drinking among students is a significant public health issue and takes a huge toll on the quality of students' lives and education. There is plenty of existing documentation on the effects of socioeconomic status on risky drinking behavior amongst college students. For example, a paper by Susan E. Collins, PhD, professor and licensed clinical psychologist, found that individuals with lower socioeconomic status as well as people of racial and ethnic minorities and homelessness experience greater alcohol-related consequences. Additionally, studies from the National Institutes of Health have found that "aspects of college life—such as unstructured time, widespread availability of alcohol, inconsistent enforcement of underage drinking laws, and limited interactions with parents and other adults" drives up rates of underage drinking, and "college students have higher binge-drinking rates and a higher incidence of driving under the influence of alcohol than their noncollege peers." Since most existing literature on underage student drinking focuses on college students, we wanted to examine the factors contributing to drinking amongst secondary school students. Our research question is as follows:

How do social indicators affect student alcohol consumption in secondary schools?

We hypothesize that factors like gender, familial status, family and school support, and other social and economic indicators will strongly influence the rates that secondary school students consume alcohol and that increased alcohol consumption is correlated with their school performance.

This data is from a Kaggle public dataset, originally sourced from the UC Irvine Machine Learning Repository. The data consists of information collected in 2008 on secondary school students from two schools in Portugal: Gabriel Pereira and Mousinho da Silveira. There are 649 observations, each one being a student, and 33 variables which cover a range of characteristics about each student's family, education, social situation, alcohol consumption, and grades in their Portuguese language class. Some key variables we will be examining are sex, male or female, Pstatus, whether their parents are together or apart, schoolsup, whether or not their school provides them extra academic support, famsup, whether or not their family provides

them extra academic support, Mjob and Fjob, the categories that their parents' jobs fall under, freetime and studytime, the amount of free time after school and amount of time spent studying on a scale of 1-5, number of absences, and amount of class failures, on a scale of 1-4. Our response variable is Walc, which is the students' average weekend alcohol consumption, on a scale of 1-5, with 5 being very high.



Methodology

We are using a multivariable linear regression to examine the effects of the predictor variables sex, Pstatus, schoolsup, famsup, Mjob, Fjob, freetime, studytime, absences, and failures on the response variable Walc, or weekend alcohol consumption. We are using a linear model rather than logistic since Walc is not binary. We started by using the stepAIC function to perform a backwards selection which allowed us to narrow down predictors with a significant effect. We then ran a VIF function to determine multicollinearity and found that our predictors were satisfactorily independent of one another. We considered interactions between studytime and absences, freetime and goout, as well as famrel and famsize. These interaction were chosen based on their real-life contexts (for instance, the effect of family size would likely change depending on familial relationship). The interactions did not significantly improve the

r-squared or rmse values as well as greatly increasing collinearity. As such, we selected Model 1 as our model.

In our model we also considered including interactions between freetime and studytime, failures and studytime, failures and absences, as well as famsup and Mjob and famsup and Fjob. In our model we used step_dummy() to create dummy variables for all categorical variables, step_interact() to create all of our interaction variables, and step_zv() to remove all variables with only one value.

Split Data into Testing and Training Data

Split Training Data in 5 Folds for Cross-Validation

```
#  5-fold cross-validation
# A tibble: 5 x 2
  splits           id
  <list>           <chr>
1 <split [388/98]> Fold1
2 <split [389/97]> Fold2
3 <split [389/97]> Fold3
4 <split [389/97]> Fold4
5 <split [389/97]> Fold5
```

StepAIC Model

term	estimate	std.error	statistic	p.value
(Intercept)	1.453	0.364	3.992	0.000
sexM	0.624	0.106	5.872	0.000
famsizeLE3	0.169	0.109	1.554	0.121
Medu	-0.218	0.066	-3.317	0.001
Fedu	0.149	0.061	2.442	0.015
Mjobhealth	0.179	0.242	0.741	0.459
Mjobother	-0.147	0.136	-1.079	0.281
Mjobservices	0.092	0.165	0.557	0.578
Mjobteacher	0.418	0.231	1.812	0.071
Fjobhealth	-0.233	0.329	-0.709	0.479
Fjobother	0.274	0.191	1.437	0.151
Fjobservices	0.419	0.205	2.046	0.041
Fjobteacher	-0.413	0.312	-1.323	0.186
reasonhome	0.011	0.127	0.085	0.932
reasonother	0.340	0.167	2.035	0.042
reasonreputation	0.202	0.133	1.519	0.130
studytime	-0.180	0.062	-2.896	0.004
nurseryyes	-0.234	0.123	-1.895	0.059
famrel	-0.172	0.052	-3.294	0.001
freetime	-0.100	0.050	-2.006	0.045
goout	0.458	0.045	10.155	0.000
health	0.108	0.035	3.090	0.002
absences	0.034	0.010	3.241	0.001

Model 1

Model Specification + Workflow 1

term	estimate	std.error	statistic	p.value
(Intercept)	1.876	0.350	5.365	0.000
studytime	-0.193	0.063	-3.080	0.002
absences	0.032	0.011	3.071	0.002
Medu	-0.240	0.066	-3.637	0.000
Fedu	0.161	0.061	2.625	0.009
famrel	-0.153	0.052	-2.922	0.004
freetime	-0.091	0.050	-1.804	0.072
goout	0.450	0.045	9.905	0.000
sex_M	0.661	0.107	6.196	0.000
Fjob_health	-0.143	0.331	-0.432	0.666
Fjob_other	0.305	0.193	1.587	0.113
Fjob_services	0.397	0.206	1.921	0.055
Fjob_teacher	-0.410	0.315	-1.303	0.193
Mjob_health	0.243	0.243	0.997	0.320
Mjob_other	-0.125	0.137	-0.909	0.364
Mjob_services	0.170	0.164	1.036	0.301
Mjob_teacher	0.485	0.232	2.091	0.037
famsize_LE3	0.156	0.110	1.420	0.156
reason_home	-0.020	0.128	-0.154	0.878
reason_other	0.301	0.168	1.793	0.074
reason_reputation	0.134	0.132	1.017	0.310
nursery_yes	-0.214	0.124	-1.719	0.086

VIF Multicollinearity Test 1

names	x
studytime	1.153
absences	1.088
Medu	2.392
Fedu	1.936
famrel	1.063
freetime	1.202
goout	1.176
sex_M	1.169
Fjob_health	1.763
Fjob_other	3.918
Fjob_services	3.695
Fjob_teacher	1.990
Mjob_health	1.695
Mjob_other	1.912
Mjob_services	1.982
Mjob_teacher	2.236
famsize_LE3	1.052
reason_home	1.232
reason_other	1.153
reason_reputation	1.269
nursery_yes	1.067

RMSE and R-Squared from 5-Fold Cross Validation 1

.metric	.estimator	mean	n	std_err	.config
rmse	standard	1.106	5	0.017	Preprocessor1_Model1
rsq	standard	0.280	5	0.015	Preprocessor1_Model1

Model 2 with Interactions

Model Specification + Workflow 2

term	estimate	std.error	statistic	p.value
(Intercept)	1.870	0.517	3.621	0.000
studytime	-0.209	0.078	-2.677	0.008
absences	0.023	0.027	0.861	0.389
Medu	-0.234	0.066	-3.525	0.000
Fedu	0.150	0.062	2.405	0.017
famrel	-0.187	0.065	-2.896	0.004
freetime	-0.049	0.130	-0.378	0.706
goout	0.500	0.133	3.769	0.000
sex_M	0.665	0.107	6.207	0.000
Fjob_health	-0.140	0.332	-0.423	0.673
Fjob_other	0.301	0.193	1.559	0.120
Fjob_services	0.392	0.207	1.893	0.059
Fjob_teacher	-0.402	0.316	-1.274	0.203
Mjob_health	0.229	0.245	0.933	0.351
Mjob_other	-0.129	0.138	-0.935	0.350
Mjob_services	0.175	0.165	1.061	0.289
Mjob_teacher	0.477	0.233	2.046	0.041
famsize_LE3	-0.265	0.450	-0.589	0.556
reason_home	-0.022	0.129	-0.171	0.865
reason_other	0.297	0.168	1.767	0.078
reason_reputation	0.131	0.133	0.985	0.325
nursery_yes	-0.211	0.125	-1.692	0.091
studytime_x_absences	0.005	0.015	0.357	0.721
freetime_x_goout	-0.014	0.038	-0.357	0.722
famsize_LE3_x_famrel	0.107	0.111	0.964	0.336

VIF Multicollinearity Test 2

names	x
studytime	1.784
absences	7.240
Medu	2.413
Fedu	1.990
famrel	1.611
freetime	7.944
goout	9.960
sex_M	1.175
Fjob_health	1.770
Fjob_other	3.931
Fjob_services	3.707
Fjob_teacher	1.996
Mjob_health	1.710
Mjob_other	1.914
Mjob_services	1.990
Mjob_teacher	2.253
famsize_LE3	17.545
reason_home	1.252
reason_other	1.155
reason_reputation	1.273
nursery_yes	1.073
studytime_x_absences	7.311
freetime_x_goout	21.846
famsize_LE3_x_famrel	18.071

RMSE and R-Squared from 5-Fold Cross Validation 2

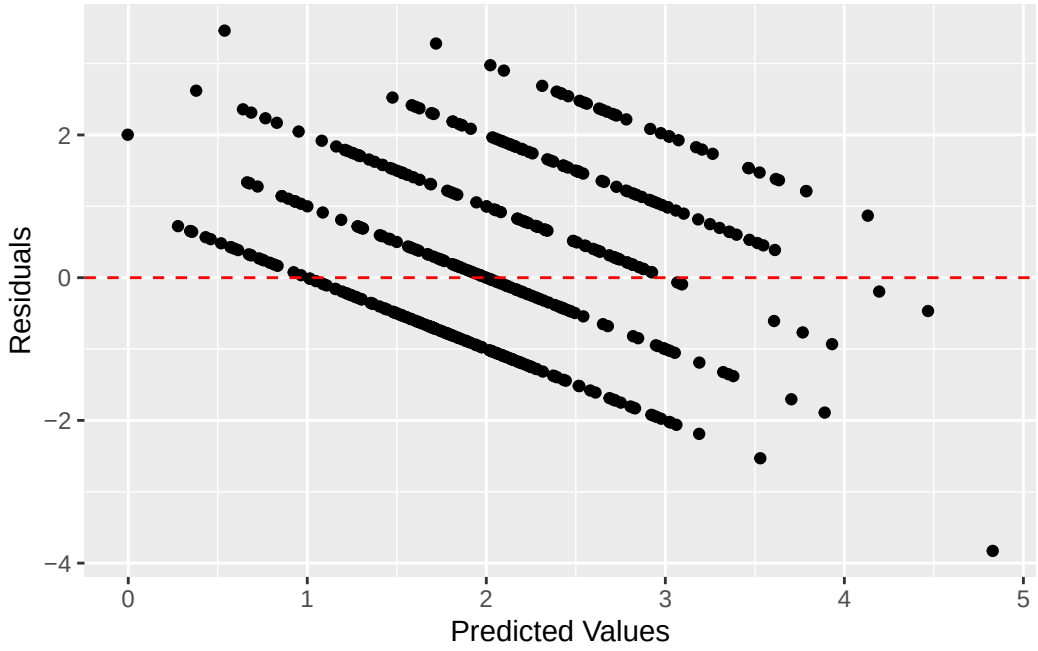
.metric	.estimator	mean	n	std_err	.config
rmse	standard	1.106	5	0.017	Preprocessor1_Model1
rsq	standard	0.280	5	0.013	Preprocessor1_Model1

Applying Model 1 to Testing Data

term	estimate	std.error	statistic	p.value
(Intercept)	1.725	0.773	2.232	0.027
studytime	-0.114	0.108	-1.056	0.293
absences	-0.008	0.022	-0.368	0.713
Medu	0.340	0.130	2.614	0.010
Fedu	-0.176	0.117	-1.508	0.134
famrel	-0.296	0.092	-3.237	0.002
freetime	0.028	0.092	0.302	0.763
goout	0.365	0.079	4.634	0.000
sex_M	0.872	0.180	4.839	0.000
Fjob_health	0.267	0.723	0.369	0.713
Fjob_other	0.066	0.470	0.141	0.888
Fjob_services	0.457	0.484	0.945	0.347
Fjob_teacher	-0.411	0.568	-0.723	0.471
Mjob_health	-0.578	0.430	-1.344	0.181
Mjob_other	-0.300	0.247	-1.213	0.227
Mjob_services	-0.716	0.305	-2.350	0.020
Mjob_teacher	-0.648	0.398	-1.627	0.106
famsize_LE3	0.080	0.187	0.429	0.669
reason_home	0.196	0.219	0.897	0.371
reason_other	0.077	0.278	0.276	0.783
reason_reputation	-0.019	0.228	-0.085	0.932
nursery_yes	-0.199	0.218	-0.911	0.364

.metric	.estimator	.estimate
rmse	standard	1.125

.metric	.estimator	.estimate
rsq	standard	0.211



Results

Our model assumes independence between predictors, linearity, and the other assumptions of a multilinear regression model. These assumptions were supported by our diagnostic tests as well as the VIF test. Our final R squared was .211 and our final RMSE was 1.125. This suggests that our model accounts for around 21% of the total variation in weekend alcohol consumption with an error of about 1.09. While the R-squared is relatively low, we believe this is still a good model given the context we are trying to predict. Predicting teen behavior is a difficult task, as there are many random factors that can attribute to decision making. Our model only covers a select few, and even with other predicting factors, teen brains usually do not align with predicable patterns.

The coefficients with significant p-values are the following:

Medu: +.340 - As each level of Mother's education increases by one unit the expected Walc score increases by .340 on average.

famrel: -.296 - As each level of quality of family relationship increases by one unit the expected Walc score decreases by -.296 on average.

goout: +.365 - As each level of going out with friends increases by one unit the expected Walc score increases by .365 on average.

sex_M: +.872 - If the individual is Male, the expected Walc score increases by .872 compared to the baseline of sex_W on average.

Mjob_services: -.716 - If the individual's Mother works in the civil services industry, the expected Walc score decreases by .716 compared to the baseline of Mjob_at_home on average.

Conclusion

From our data analysis, we conclude that family relationships has the biggest impact on weekend alcohol consumption. Extrapolating from this, we emphasize the importance of strong family dynamics for school aged children. This drives us to consider more research into what makes children feel supported by their families and how parents' employment impacts daily decisions. We also found it interesting that men are more likely to participate in alcohol consumption at school age rather than women. We believe that may have something to local gender norms and allows for further research into young male substance abuse and mental health.

Our model is significantly limited by the data collection. The data comes from two schools in Portugal, making it unrealistic to apply to American teens, especially given the different alcohol laws in United States. Portugal also has different family dynamics and values than the US. Our data collection also is self reported on a scale of 1 to 5. Using a 1-5 scale limits our numerical analysis because every evaluation is relative to each participant's understanding of their own situation. This is why we also saw a trend towards the middle, since most participants are likely to pick a less extreme answer in a self reported context. If we were to replicate this analysis, we would seek to find data from the United States that has similar predictor variables, but ones that a measured quantitatively instead of on a 1-5 scale. An example of this would be estimating how many drinks consumed on the weekend rather than self reporting on the scale. We are satisfied with the categorical variables.