Alcohol Consumption in Schools

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

2023-11-14

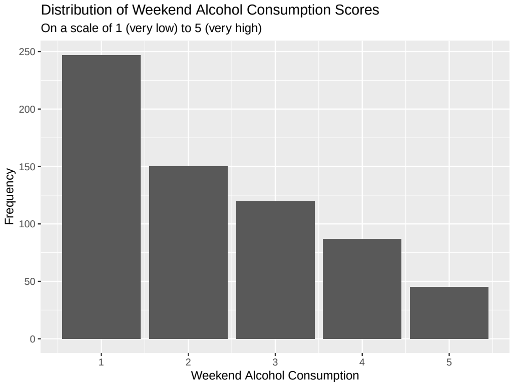
Introduction and data

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:

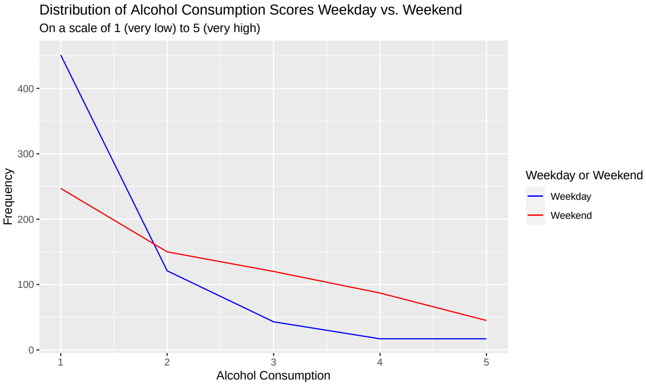
How do social indicators affect student alcohol consumption in secondary schools?

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.

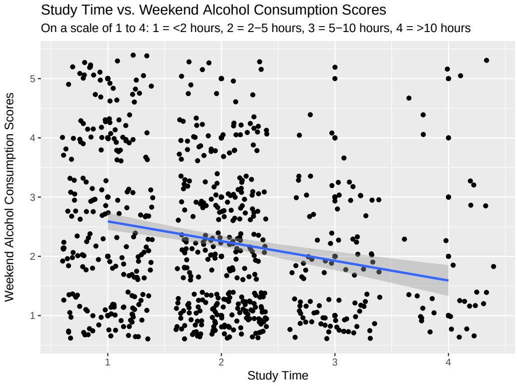
This data is from a Kaggle public dataset, originally sourced from the UC Irvine Machine Learning Repository. The data consists of information collected in 2008 on secondary school students from two schools in Portugal: Gabriel Pereira and Mousinho da Silveira. There are 649 observations, each one being 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.


Distribution of Weekend Alcohol Consumption Scores

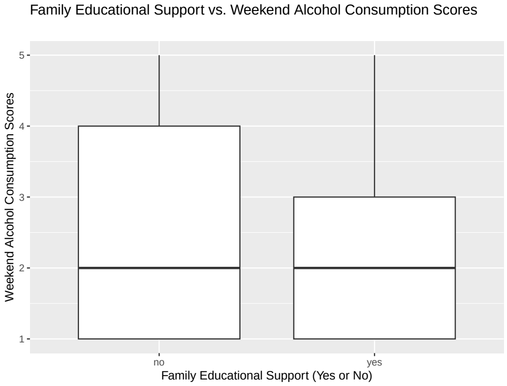
On a scale of 1 (very low) to 5 (very high)


Distribution of Alcohol Consumption Scores Weekday vs. Weekend

On a scale of 1 (very low) to 5 (very high)


Study Time vs. Weekend Alcohol Consumption Scores

On a scale of 1 to 4: 1 = <2 hours, 2 = 2−5 hours, 3 = 5−10 hours, 4 = >10 hours


Family Educational Support vs. Weekend Alcohol Consumption Scores

Distirbution of Response Variable

minimum	q1	median	mean	q3	maximum
1	1	2	2.280431	3	5

Walc	count	Percentage
1	247	38.06%
2	150	23.11%
3	120	18.49%
4	87	13.41%
5	45	6.93%

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. With many behavioral or otherwise psychological models, the main focus is to identify the greatest effects rather than any effects. As such, because we weren't sure which predictors would give us significant effects, we started by using the stepAIC function to perform a backwards selection on the full model which allowed us to narrow down the predictors with significant effects. The backwards selection process took the full model and removed individual variables until the AIC no longer decreased. 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.

The StepAIC and VIF steps can be found in the appendix.

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
```

Model 1

Model Specification + Workflow 1

term	estimate	std.error	statistic	p.value
(Intercept)	1.583	0.359	4.404	0.000
studytime	-0.180	0.062	-2.896	0.004
absences	0.034	0.010	3.241	0.001
Medu	-0.218	0.066	-3.317	0.001
Fedu	0.149	0.061	2.442	0.015
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
sex_M	0.624	0.106	5.872	0.000
Fjob_health	-0.233	0.329	-0.709	0.479
Fjob_other	0.274	0.191	1.437	0.151
Fjob_services	0.419	0.205	2.046	0.041
Fjob_teacher	-0.413	0.312	-1.323	0.186
Mjob_health	0.179	0.242	0.741	0.459
Mjob_other	-0.147	0.136	-1.079	0.281
Mjob_services	0.092	0.165	0.557	0.578
Mjob_teacher	0.418	0.231	1.812	0.071
famsize_LE3	0.169	0.109	1.554	0.121
reason_home	0.011	0.127	0.085	0.932
reason_other	0.340	0.167	2.035	0.042
reason_reputation	0.202	0.133	1.519	0.130
nursery_yes	-0.234	0.123	-1.895	0.059

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

.metric	.estimator	mean	n	std_err	.config
rmse	standard	1.103	5	0.014	Preprocessor1_Model1
rsq	standard	0.286	5	0.014	Preprocessor1_Model1

Model 2 with Interactions

Model Specification + Workflow 2

term	estimate	std.error	statistic	p.value
(Intercept)	1.651	0.517	3.195	0.001
studytime	-0.191	0.078	-2.456	0.014
absences	0.028	0.027	1.032	0.303
Medu	-0.211	0.066	-3.187	0.002
Fedu	0.137	0.062	2.219	0.027
famrel	-0.213	0.065	-3.297	0.001
freetime	-0.076	0.129	-0.592	0.554
goout	0.490	0.131	3.731	0.000
health	0.110	0.035	3.112	0.002
sex_M	0.627	0.107	5.871	0.000
Fjob_health	-0.228	0.330	-0.689	0.491
Fjob_other	0.272	0.192	1.419	0.157
Fjob_services	0.417	0.205	2.031	0.043
Fjob_teacher	-0.405	0.313	-1.296	0.196
Mjob_health	0.160	0.244	0.656	0.512
Mjob_other	-0.152	0.136	-1.113	0.266
Mjob_services	0.093	0.165	0.565	0.572
Mjob_teacher	0.405	0.232	1.746	0.081
famsize_LE3	-0.318	0.446	-0.713	0.476
reason_home	0.011	0.128	0.089	0.929
reason_other	0.335	0.167	2.006	0.045
reason_reputation	0.200	0.133	1.501	0.134
nursery_yes	-0.234	0.124	-1.885	0.060
studytime_x_absences	0.004	0.015	0.242	0.809
freetime_x_goout	-0.008	0.038	-0.211	0.833
famsize_LE3_x_famrel	0.124	0.110	1.127	0.260

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

.metric	.estimator	mean	n	std_err	.config
rmse	standard	1.104	5	0.014	Preprocessor1_Model1
rsq	standard	0.285	5	0.013	Preprocessor1_Model1

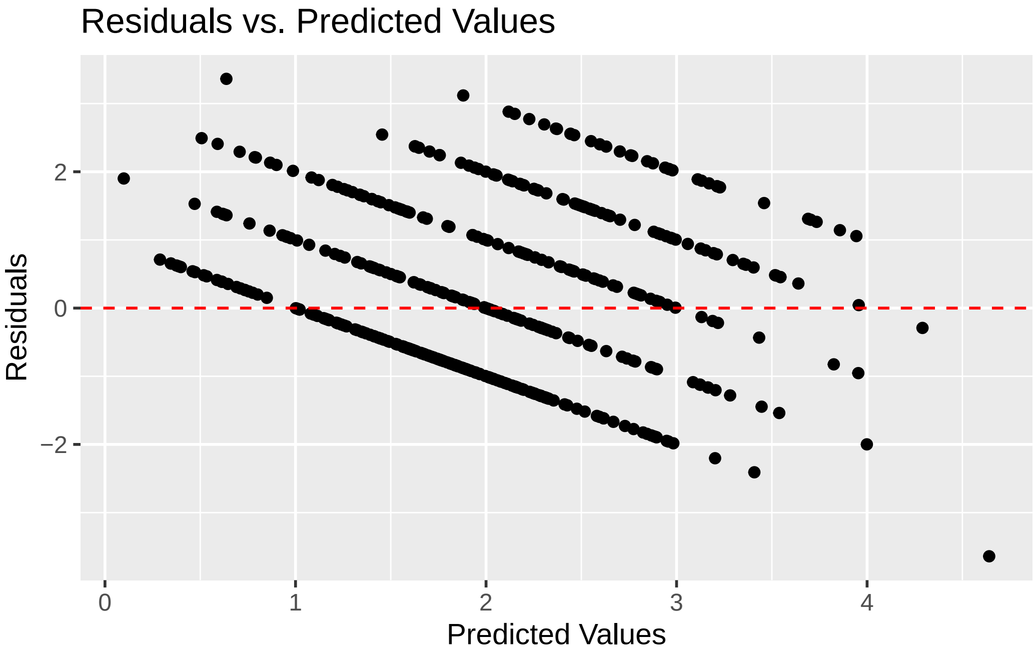
Applying Model 1 to Testing Data

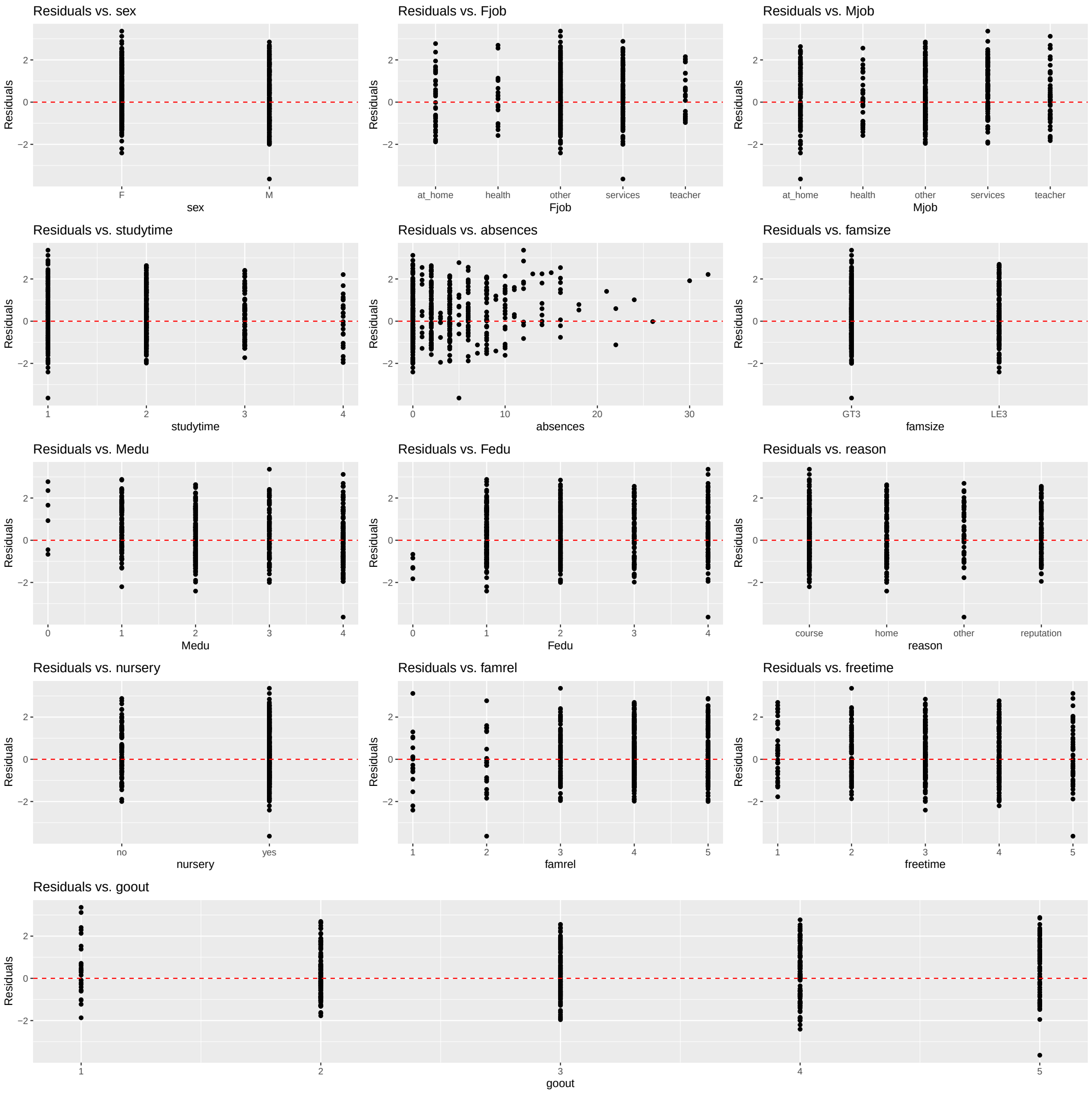
term	estimate	std.error	statistic	p.value
(Intercept)	1.455	0.797	1.825	0.070
studytime	-0.127	0.108	-1.172	0.243
absences	-0.011	0.022	-0.519	0.605
Medu	0.345	0.130	2.657	0.009
Fedu	-0.173	0.117	-1.483	0.140
famrel	-0.307	0.092	-3.347	0.001
freetime	0.022	0.092	0.241	0.810
goout	0.381	0.079	4.794	0.000
health	0.083	0.062	1.330	0.186
sex_M	0.838	0.181	4.616	0.000
Fjob_health	0.138	0.727	0.189	0.850
Fjob_other	0.041	0.469	0.087	0.931
Fjob_services	0.425	0.483	0.880	0.380
Fjob_teacher	-0.454	0.568	-0.800	0.425
Mjob_health	-0.640	0.431	-1.483	0.140
Mjob_other	-0.319	0.247	-1.290	0.199
Mjob_services	-0.737	0.304	-2.420	0.017
Mjob_teacher	-0.654	0.397	-1.647	0.102
famsize_LE3	0.063	0.187	0.339	0.735
reason_home	0.214	0.218	0.980	0.329
reason_other	0.055	0.277	0.198	0.843
reason_reputation	0.009	0.229	0.039	0.969
nursery_yes	-0.155	0.220	-0.706	0.481

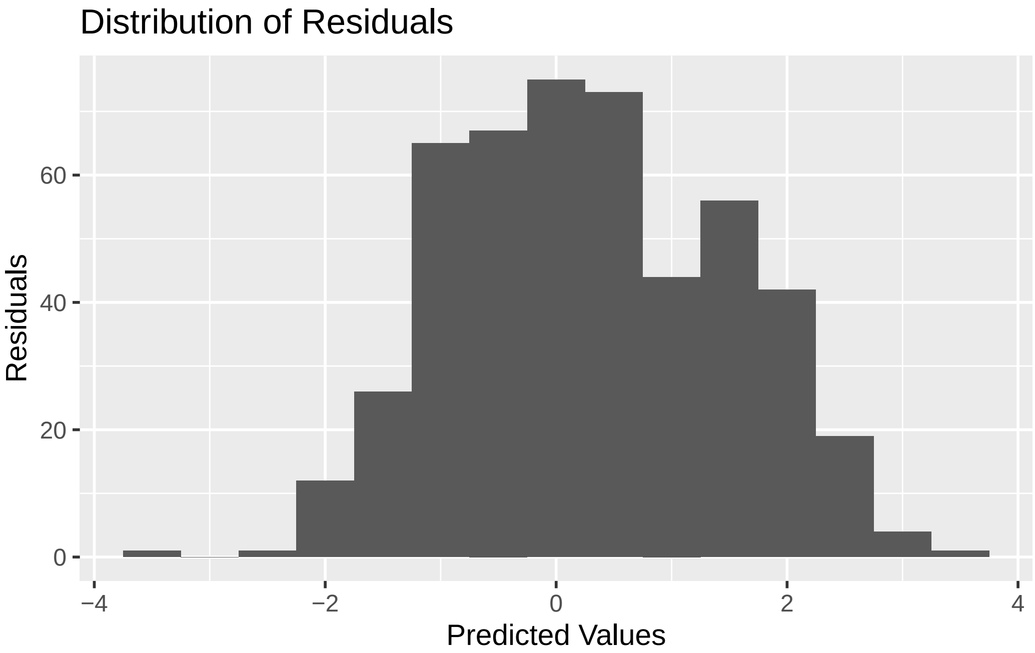
.metric	.estimator	.estimate
rmse	standard	1.121

.metric	.estimator	.estimate
rsq	standard	0.224

Checking Conditions for Selected Model




Residuals vs. sex
Residuals
2
0
−2
F
M
sex
Residuals vs. Fjob
at_home
health
other
services
teacher
Fjob
Residuals vs. Mjob
at_home
health
other
services
teacher
Mjob
Residuals vs. studytime
1
2
3
4
studytime
Residuals vs. absences
0
10
20
30
absences
Residuals vs. famsize
GT3
LE3
famsize
Residuals vs. Medu
0
1
2
3
4
Medu
Residuals vs. Fedu
Fedu
Residuals vs. reason
course
home
other
reputation
reason
Residuals vs. nursery
no
yes
nursery
Residuals vs. famrel
1
2
3
4
5
famrel
Residuals vs. freetime
freetime
Residuals vs. goout
goout



Linearity: Linearity is met because there are no discernible patterns in the residuals plot of the predicted values and all of the individual predictors; the points are scattered relatively randomly.

Constant Variance: Constant Variance is met because in the bulk of the data, the vertical spread is relatively constant throughout; there does not appear to be a fan shape as you move from left to right.

Normality: Normality is satisfied because the distribution of residuals is approximately unimodal and symmetric.

Independence: Independence is satisfied because the data is a random sample, and we can reasonably assume that the data for one student does not effect the data for the other students.

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.

Appendix

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

VIF Multicollinearity Test 1

names	x
studytime	1.158
absences	1.090
Medu	2.420
Fedu	1.944
famrel	1.078
freetime	1.207
goout	1.180
health	1.145
sex_M	1.184
Fjob_health	1.777
Fjob_other	3.929
Fjob_services	3.700
Fjob_teacher	1.990
Mjob_health	1.707
Mjob_other	1.918
Mjob_services	2.030
Mjob_teacher	2.255
famsize_LE3	1.054
reason_home	1.239
reason_other	1.160
reason_reputation	1.304
nursery_yes	1.070

VIF Multicollinearity Test 2

names	x
studytime	1.795
absences	7.260
Medu	2.444
Fedu	1.998
famrel	1.637
freetime	7.980
goout	9.966
health	1.152
sex_M	1.190
Fjob_health	1.783
Fjob_other	3.941
Fjob_services	3.712
Fjob_teacher	1.996
Mjob_health	1.724
Mjob_other	1.920
Mjob_services	2.041
Mjob_teacher	2.276
famsize_LE3	17.570
reason_home	1.261
reason_other	1.161
reason_reputation	1.310
nursery_yes	1.077
studytime_x_absences	7.321
freetime_x_goout	21.896
famsize_LE3_x_famrel	18.116