The Effect of Alcohol Consumption on Student Grades

Vanessa Tang

12/10/2019

Summary

What factors influence student grades? Social, demographic, family, and academic information can provide insight on student performance in school. For this analysis, we aim to identify factors that influence student grades in addition to how alcohol consumption specifically impacts grades. To do so, important and signficant covariates were identified, and a proportional odds model was fit to predict students' grades. Based on the results of this proportional odds model, we found that number of past class failures and planning on pursuing higher education are the two strongest predictors of final grade. Furthermore, consuming alcohol is likely to decrease the odds of a student receiving higher grades. Overall, this study aims to identify factors that influence student academic performance.

Introduction

While there are a wide variety of factors that may influence a student's academic performance, this analysis will only focus on a few select predictors. In order to help determine what affects student grades, we will be aiming to answer three main questions.

1. What factors are the strongest predictors of overall student grades?
2. How does alcohol consumption affect overall student grades?
3. What other factors affect overall student grades?

More specifically, given this data including various social, demographic, family, and academic information for 10th through 12th graders, we aim to identify specific variables that influence student grades the most. For example, how do internet access, mother's occupation, father's education, extra tutoring sessions, or hours of free time impact a student's grades? How do extracurricular activities play a role? It is often considered that "going out" or excessive social time can decrease student's grades, but based on this data, is this actually true? This study aims to answer these questions, identifying factors that not only impact student grades but also surprisingly may not affect student grades.

Data

The dataset was obtained from Kaggle and was previously used in a paper by Paulo Cortez and Alice Silva titled "Using Data Mining to Predict Secondary School Student Performance." While their paper implements data mining models using all available data including previous periodic grades, this study focuses more on identifying specific factors that affect overall grade. The dataset includes demographic, social, family, and academic information and was gathered through a combination of academic records and self-reported surveys from 10th through 12th grade students (secondary school) from two public schools in Portugal for the 2005-2006 academic year. Notably, the legal drinking age in Portugal during 2005 and 2006 was 16, and 850 of the 1,044 students are 16 years of age or older. A summary of the variables available in the dataset are listed below, and a full data dictionary can be found in the appendix.

Summary of Variables:

Demographic	Social	Family	Academic
sex	activities	famsize	school
age	romantic	Pstatus	studytime

Demographic	Social	Family	Academic
address	goout	Medu	failures
traveltime	freetime	Fedu	schoolsup
internet	Dalc	famsup	paid
health	Walc	famrel	nursery
	Mjob	higher	
	Fjob	absences	
	guardian	subject	
		reason	
		G1	
		G2	
		G3	

The original dataset contains 33 variables for 1,044 students. There are 53 students who have a final grade `G3` of 0. Many of these students have data for the second period grade `G2`, but all have a grade for the first period grade `G1`. This pattern indicates that these students who have grades of 0 for `G3` may not have actually finished the course. Because course incompletion is different from actually receiving a 0 in the course, these students will be dropped from the dataset. No other missing variables are present. After dropping the students who have zeros for `G3`, the dataset has 991 students.

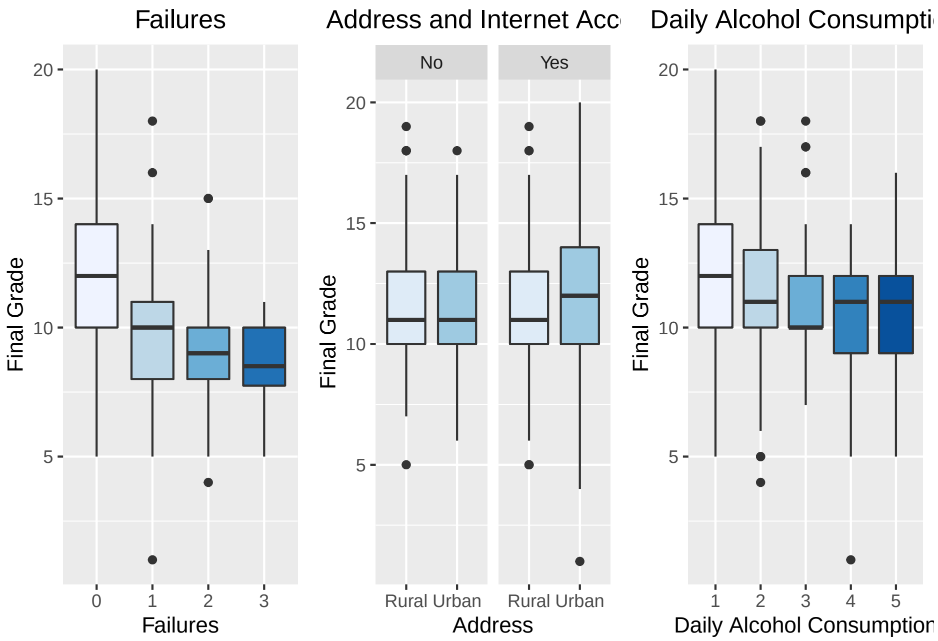
Final grades `G3` were binned into 5 levels. The observed mean (11.95) and median (12) of final grades `G3` were used to help generate bins. Thus, these bins are more representative of students' grades relative to the observed mean. A new variable `grade` was added according to the following table which also shows the counts for each binned grade level:

`G3` range	`grade`	count
20 - 16	excellent	122
15 - 13	very good	285
12 - 11	average	254
10 - 9	sufficient	216
9 - 0	weak	114

For exploratory data analysis, chi-squared tests were conducted between `grade` and all predictors. At a significance of 0.05, there was a significant difference between `grade` and the following predictors: `school`, `address`, `studytime`, `schoolsup`, `paid`, `nursery`, `higher`, `internet`, `freetime`, `goout`, `health`, `subject`, `Mjob`, and `reason`. This provides an indication that `grade` differs among the different levels of each of the previously mentioned predictors.

However, some variables could not be used for chi-squared tests due to lack of data in specific combinations of variables and `grade` levels. Further model testing will account for this through deviance tests between nested models.

To further examine potential relationship between the various predictors and `grade`, plots were generated using the unbinned grade variable `G3` of final grades on a scale of 0 (low) to 20 (high). From the following plots of `failures`, `address`, `internet`, and `Dalc` against `G3`, it is evident that generally, students who had more past class failures tended to have lower final grades. Interestingly, those who live in urban areas and have internet access at home also tended to have higher grades than others. This indicates a possible interaction term between `address` and `internet`. Lastly, in the far right plot of `Dalc` against `G3`, there seems to be a very slight decrease in grade for increasing levels of `Dalc`, but further modeling is necessary to clarify this relationship. Other EDA plots with `G3` as the dependent variable can be found in the appendix.



Model

For this analysis, a proportional odds model was chosen because the response variable `grade` has natural ordering for the five levels. Based on the significant variables mentioned in EDA and other variables of interest, predictors were added and removed through many deviance tests comparing nested models. Stepwise selection using AIC was also used for variable selection. Lastly, interaction terms were tested using deviance tests. Variables and interactions were selected to remain in the final model if the p-value of the deviance test was below 0.05. Using this method, the final model is as follows:

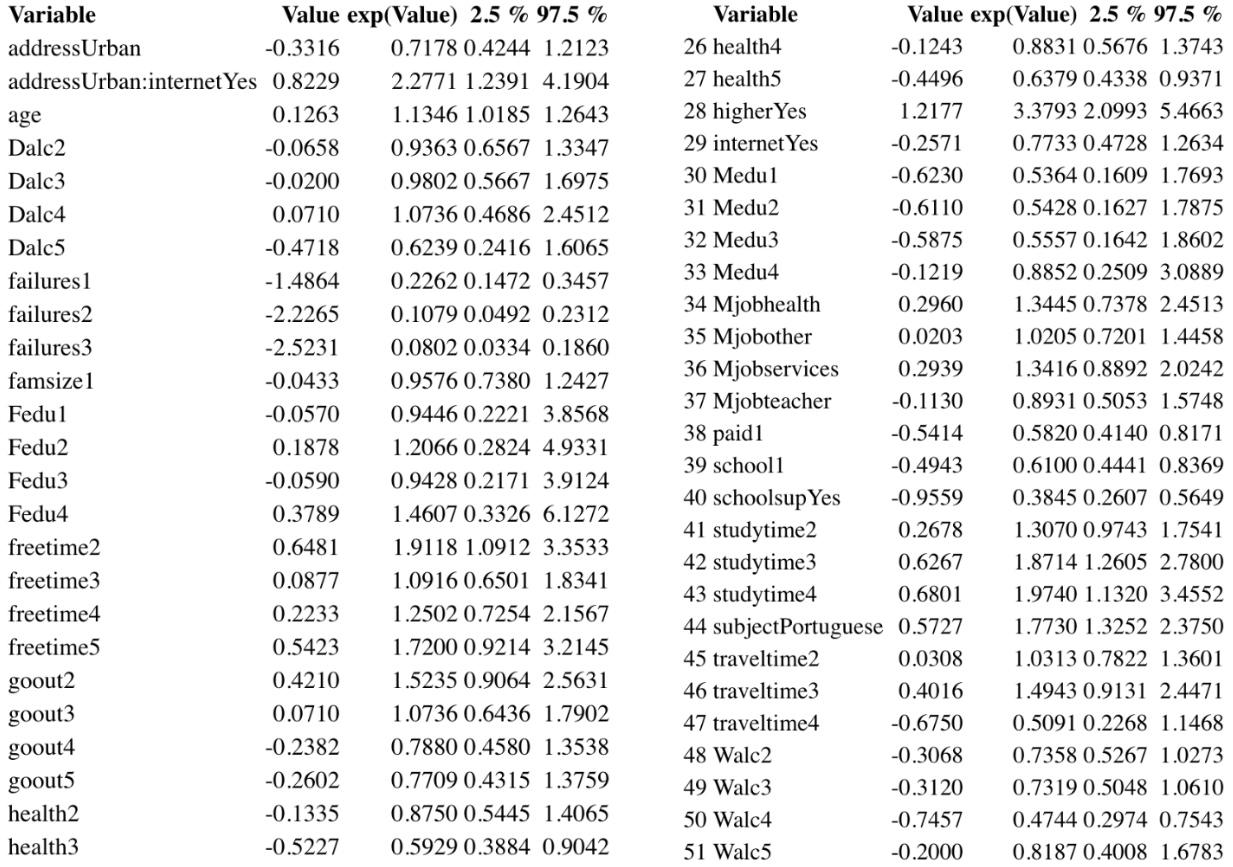
$$log(\frac{Pr[grade_i \leq j|x_i]}{Pr[grade_i > j|x_i]}) = x_i\beta, j = 1, ..., 4$$

where x contains `failures`, `Medu`, `Dalc`, `Walc`, `higher`, `subject`, `studytime`, `address`, `freetime`, `internet`, `Fedu`, `age`, `Mjob`, `traveltime`, `famsize`, `schoolsup`, `goout`, `school`, `paid`, `health`, and an interaction between `address` and `internet`.

This model has an AIC of 2770.58 and overall accuracy of 40.06%. While this accuracy may seem low, this only portrays the explicitly correct predictions of `grade` and does not truly account for the ordered nature of the levels. The confusion matrix is as follows where rows are predicted levels and columns are actual levels. In this instance, the sensitivity is the count of correct predictions divided by the actual count for that variable (column total). Sensitivity helps gauge how well the model predicted within each observed bin level for `grade`. Based on these results, we can see that the model does a decent job at predicting "very_good" but does not perform as well in the extremes of "weak" and "excellent".

Prediction	weak	sufficient	average	very_good	excellent	predicted totals
weak	25	28	4	2	0	59
sufficient	45	77	50	22	5	199
average	33	62	93	60	10	258
very_good	11	48	106	193	98	456
excellent	0	1	1	8	9	19
actual totals	114	216	254	285	122	
sensitivity	0.22	0.36	0.37	0.68	0.074	

The resulting model coefficients ("Value"), exponentiated coefficients ("exp(Value)"), and exponentiated 95% confidence intervals are shown below. Full model results including standard errors and t-values can be found in the appendix.



Variable	Value	exp(Value)	2.5 %	97.5 %
addressUrban	-0.3316	0.7178	0.4244	1.2123
addressUrban:internetYes	0.8229	2.2771	1.2391	4.1904
age	0.1263	1.1346	1.0185	1.2643
Dalc2	-0.0658	0.9363	0.6567	1.3347
Dalc3	-0.0200	0.9802	0.5667	1.6975
Dalc4	0.0710	1.0736	0.4686	2.4512
Dalc5	-0.4718	0.6239	0.2416	1.6065
failures1	-1.4864	0.2262	0.1472	0.3457
failures2	-2.2265	0.1079	0.0492	0.2312
failures3	-2.5231	0.0802	0.0334	0.1860
famsize1	-0.0433	0.9576	0.7380	1.2427
Fedu1	-0.0570	0.9446	0.2221	3.8568
Fedu2	0.1878	1.2066	0.2824	4.9331
Fedu3	-0.0590	0.9428	0.2171	3.9124
Fedu4	0.3789	1.4607	0.3326	6.1272
freetime2	0.6481	1.9118	1.0912	3.3533
freetime3	0.0877	1.0916	0.6501	1.8341
freetime4	0.2233	1.2502	0.7254	2.1567
freetime5	0.5423	1.7200	0.9214	3.2145
goout2	0.4210	1.5235	0.9064	2.5631
goout3	0.0710	1.0736	0.6436	1.7902
goout4	-0.2382	0.7880	0.4580	1.3538
goout5	-0.2602	0.7709	0.4315	1.3759
health2	-0.1335	0.8750	0.5445	1.4065
health3	-0.5227	0.5929	0.3884	0.9042
26 health4	-0.1243	0.8831	0.5676	1.3743
27 health5	-0.4496	0.6379	0.4338	0.9371
28 higherYes	1.2177	3.3793	2.0993	5.4663
29 internetYes	-0.2571	0.7733	0.4728	1.2634
30 Medu1	-0.6230	0.5364	0.1609	1.7693
31 Medu2	-0.6110	0.5428	0.1627	1.7875
32 Medu3	-0.5875	0.5557	0.1642	1.8602
33 Medu4	-0.1219	0.8852	0.2509	3.0889
34 Mjobhealth	0.2960	1.3445	0.7378	2.4513
35 Mjobother	0.0203	1.0205	0.7201	1.4458
36 Mjobservices	0.2939	1.3416	0.8892	2.0242
37 Mjobteacher	-0.1130	0.8931	0.5053	1.5748
38 paid1	-0.5414	0.5820	0.4140	0.8171
39 school1	-0.4943	0.6100	0.4441	0.8369
40 schoolsupYes	-0.9559	0.3845	0.2607	0.5649
41 studytime2	0.2678	1.3070	0.9743	1.7541
42 studytime3	0.6267	1.8714	1.2605	2.7800
43 studytime4	0.6801	1.9740	1.1320	3.4552
44 subjectPortuguese	0.5727	1.7730	1.3252	2.3750
45 traveltime2	0.0308	1.0313	0.7822	1.3601
46 traveltime3	0.4016	1.4943	0.9131	2.4471
47 traveltime4	-0.6750	0.5091	0.2268	1.1468
48 Walc2	-0.3068	0.7358	0.5267	1.0273
49 Walc3	-0.3120	0.7319	0.5048	1.0610
50 Walc4	-0.7457	0.4744	0.2974	0.7543
51 Walc5	-0.2000	0.8187	0.4008	1.6783

1. What factors are the strongest predictors of overall student grades?

Based on the t-values of the various coefficients in this model, the strongest predictor of `grade` is the number of past class failures. For any fixed grade level, the estimated odds that a student with 1 failure is in the "weak" direction rather than the "excellent" direction is 0.23 times the estimated odds for a student no past class failures. Second to `failures`, `higher` is also a strong predictor of `grade`. For any fixed grade level, the odds that a student who wants to pursue higher education scores in the "excellent" direction is 3.38 times that of a student who is not planning on pursuing higher education.

2. How does alcohol consumption affect overall student grades?

Based on the 95% confidence intervals generated from this model, all 4 levels of daily and 3 of the 4 levels of weekly alcohol consumption contain 0, indicating that they are significant relative to baseline where daily and weekly alcohol consumption are at level 1 (low). For daily alcohol consumption, for any fixed grade level, the estimated odds that a student whose daily alcohol consumption is at levels 2, 3, and 5 is in the "weak" direction rather than the "excellent" direction are 0.94, 0.98, and 0.62 times the estimated odds for a student whose daily alcohol consumption is reported to be at level 1. However, because the coefficient for daily alcohol consumption at level 4 is positive, the estimated odds that these students are in the "excellent" direction rather than the "weak" direction is 1.07 times those at a daily alcohol consumption of level 1. This result is not consistent with the results for other levels of `Dalc` and may be due to the self-reported nature of this data, as very few students report daily alcohol consumption at the higher levels.

For weekly alcohol consumption, for any fixed grade level, the estimated odds that a student whose weekly alcohol consumption is at levels 2, 3, 4, and 5 is in the "weak" direction rather than the "excellent" direction are 0.74, 0.73, 0.47, and 0.82 times the estimated odds for a student whose weekly alcohol consumption is at the baseline level 1. Based on these results of daily and weekly alcohol consumption, it seems that generally, those who consume more alcohol have higher odds of receiving weaker grades. However, each increasing level of alcohol consumption (relative to level 1) does not have a clear, increasing pattern in odds of having weaker grades. Overall, inconsistence in the results for the various levels of both `Dalc` and `Walc` make it difficult to conclusively determine the effect of alcohol consumption on student grades.

3. What other factors affect overall student grades?

While there are several factors that affect overall grade, one interesting result from this model is that for any fixed grade level, the estimated odds that a student living in an urban address with internet access at home is in the "excellent" direction is 2.28 times that of a student living in a rural area with no internet access at home. Second, for any fixed grade level, the estimated odds that a student whose weekly study hours are at the 2 (2 - 5 hrs), 3 (5 - 10 hrs), and 4 (> 10 hrs) levels are in the "excellent" direction are 1.31, 1.87, and 1.97 times the estimated odds for baseline students at level 1 (<2 hours) of weekly study time. While not necessarily surprising, this confirms that the more a student studies, the higher their odds are of receiving grades closer to the "excellent" level rather than the "weak" level. Lastly, based on the positive coefficient of `freetime`, relative to baseline students with after school freetime of level 1 (1 = very low, 5 = very high), students with more freetime have higher odds of receiving higher grades. These are just some of the most interesting results from this model, and more details on odds and confidence intervals for each variable can be found in the results table above.

Conclusion

Based on this analysis of the various factors that affect a student's academic performance, we found that past class failures and planning on pursuing higher education are the two most important predictors of student grades given this data. Furthermore, in general, those who consumed higher weekly levels of alcohol consumption had higher odds of having lower grades. This analysis also confirms that studying more tends to increase a student's odds of receiving higher grades. Lastly, living in an urban area with internet access at home increased the odds of receiving higher grades relative to those in rural areas with no internet access at home. Overall, these findings are important because they emphasize the importance of not failing classes, studying more, planning on pursuing higher education, and decreasing student alcohol consumption. Possibly most notable, however, is the interaction among at-home internet access, urban living, and student grades. At a greater level, this hints at a possible discrepancy among rural and urban populations and access to technology that could inhibit a student's academic performance. While the present study does not touch on this subject, it would be interesting to further research specifically how rural and urban populations differ in terms of technological access and how it may affect academics.

While this analysis aimed to identify important factors in high school students' academic performance, one major limitation is the lack of other data or variables. For example, household income could also be a strong predictor of academic performance, but in this dataset, this information was not available. Second, this data only includes students in grades 10 through 12 in Portugal, representing a small portion of all students. Other environmental or school system-based factors may contribute to a student's academic performance. Most notably, however, is that this data was gathered through surveys and is thus self-reported. It is possible that this data is not truly representative of students' habits, as students may not have been truthful in reporting factors such as alcohol consumption, study time, and freetime. Future work should aim to gather more data from a larger portion of the population and include other possibly important variables such as income. Ideally, data would be gathered in a different way, possibly by tracking students rather than relying on students' self-reported survey results.

Despite these limitations and the need for future work, this study is successful in identifying important areas that may affect student grades. While academic performance can be dependent on a wide variety of factors, this study specifically highlights the importance of passing classes, striving for higher education, and dedicating more hours to studying.

Appendix

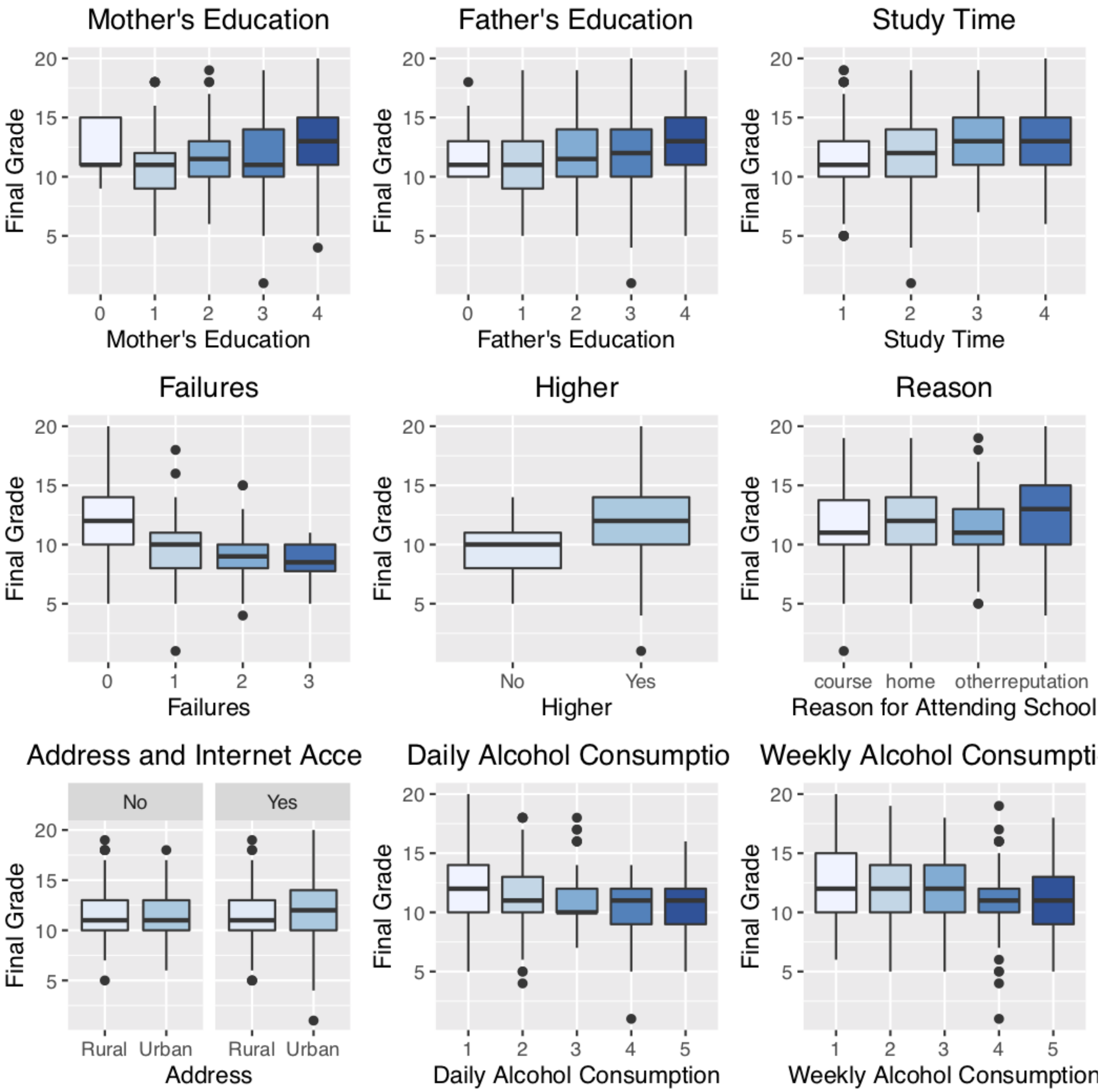
GitHub Repository

https://github.com/vmtang11/student_grades/

Data Dictionary

Attribute	Description
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: 0 - none, 1 - to 4th grade, 2 – 5th to 9th grade, 3 – 10th to 12th grade or 4 – higher education)
Mjob	mother's job (nominal: teacher, health care related, civil services, at home or other)
Fedu	father's education (numeric: 0 - none, 1 - to 4th grade, 2 – 5th to 9th grade, 3 – 10th to 12th grade or 4 – higher education)
Fjob	father's job (nominal: teacher, health care related, civil services, at home or other)
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time(numeric: 1 – <15min., 2 – 15 to 30 min., 3 – 30min. to 1 hour or 4 – > 1 hour).
studytime	weekly study time (numeric: 1 – <2hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paid	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)
subject	class identifier (math or portuguese)

EDA Plots



Model Results

$log(\frac{Pr[grade_i \leq j|x_i]}{Pr[grade_i > j|x_i]}) = \beta_{0j} + \beta_1 * failures_{i1} + \beta_2 * Medu_{i2} + \beta_3 * Dalc_{i3} + \beta_4 * higher_{i4} + \beta_5 * subject_{i5} + \beta_6 * studytime_{i6} + \beta_7 * address_{i7} + \beta_8 * freetime_{i8} + \beta_9 * internet_{i9} + \beta_{10} * Fedu_{i10} + \beta_{11} * age_{i11} + \beta_{12} * Walc_{i12} + \beta_{13} * Mjob_{i13} + \beta_{14} * traveltime_{i14} + \beta_{15} * famsize_{i15} + \beta_{16} * schoolsup_{i16} + \beta_{17} * goout_{i17} + \beta_{18} * school_{i18} + \beta_{19} * paid_{i19} + \beta_{20} * health_{i20} + \beta_{21} * address : internet_{i21}, j = 1, ..., 4$

Re-fitting to get Hessian

Call: polr(formula = grade ~ failures + Medu + Dalc + higher + subject + studytime + address + freetime + internet + Fedu + age + Walc + Mjob + traveltime + famsize + schoolsup + goout + school + paid + health + address:internet, data = students)

Table 5: Coeficients

	Value	Std. Error	t value
failures1	-1.486	0.2176	-6.83
failures2	-2.226	0.3932	-5.663
failures3	-2.523	0.4358	-5.79
Medu1	-0.623	0.6068	-1.027
Medu2	-0.611	0.6065	-1.007
Medu3	-0.5875	0.6146	-0.9559
Medu4	-0.1219	0.6361	-0.1917
Dalc2	-0.06582	0.1808	-0.364
Dalc3	-0.02004	0.2796	-0.07169
Dalc4	0.07097	0.4208	0.1687
Dalc5	-0.4718	0.4821	-0.9787
higherYes	1.218	0.2439	4.993
subjectPortuguese	0.5727	0.1488	3.849
studytime2	0.2678	0.15	1.785
studytime3	0.6267	0.2017	3.107
studytime4	0.6801	0.2843	2.392
addressUrban	-0.3316	0.2676	-1.239
freetime2	0.6481	0.2862	2.265
freetime3	0.08768	0.2643	0.3317
freetime4	0.2233	0.2777	0.8042
freetime5	0.5423	0.3185	1.703
internetYes	-0.2571	0.2506	-1.026
Fedu1	-0.05705	0.7248	-0.0787
Fedu2	0.1878	0.7265	0.2585
Fedu3	-0.05895	0.7348	-0.08023
Fedu4	0.3789	0.7404	0.5118
age	0.1263	0.05511	2.291
Walc2	-0.3068	0.1704	-1.801
Walc3	-0.312	0.1894	-1.648
Walc4	-0.7457	0.2373	-3.142
Walc5	-0.2	0.3649	-0.5481
Mjobhealth	0.296	0.3061	0.9671
Mjobother	0.02029	0.1777	0.1142
Mjobservices	0.2939	0.2098	1.401
Mjobteacher	-0.113	0.2899	-0.39
traveltime2	0.03082	0.1411	0.2185
traveltime3	0.4016	0.2512	1.599
traveltime4	-0.675	0.412	-1.639
famsize1	-0.04332	0.1329	-0.326
schoolsupYes	-0.9559	0.1972	-4.848
goout2	0.421	0.2649	1.589
goout3	0.07103	0.2607	0.2725
goout4	-0.2382	0.2762	-0.8624
goout5	-0.2602	0.2956	-0.8802
school1	-0.4943	0.1616	-3.059
paid1	-0.5414	0.1734	-3.122
health2	-0.1335	0.242	-0.5516
health3	-0.5227	0.2155	-2.425
health4	-0.1243	0.2255	-0.5512
health5	-0.4496	0.1964	-2.289

	Value	Std. Error	t value
addressUrban:internetYes	0.8229	0.3107	2.649

Table 6: Intercepts

	Value	Std. Error	t value
weak|sufficient	0.07246	1.415	0.05121
sufficient|average	1.835	1.414	1.298
average|very_good	3.263	1.416	2.305
very_good|excellent	5.217	1.423	3.666

Residual Deviance: 2660.583

AIC: 2770.583