Part II : Nature or Nurture ? Examining the factors that influence academic achievement in Adolescents

Multiple Regression Model

[redacted]

[redacted]

Introduction

The initial exploration in Part I of this assignment was focused on identifying factors, which influence academic performance in adolescents. Understanding the factors behind academic success will help teachers and departments of education better understand where to divert more resources and at a government level it will help determine policy decisions. This paper will continue and build upon that analysis identifying additional factors, which contribute towards academic performance, and then combining all investigated factors to build a multiple linear regression model capable of predicting a student's academic performance.

The initial analysis found that factors such as Mother's Education, Household Income and Athleticism all had a small but significant contribution toward a students academic performance. Additional research on the topic has suggested that gender also plays a key role in academic success amongst high school students. *(Sayid Dabbagh Ghazvinia, Milad Khajehpour, 2011)* investigated the gender differences affecting academic performance. They found that girls made greater use of attitude, motivation, time management, anxiety and self-testing strategies and tended to perform better in Literature. The boys tended to use more concentration, information processing and selecting main idea strategies and tended to perform better in Maths.

Other studies have noted that ethnicity also plays an important role in academic achievement. *(Palardy, Rumberger, Butler, 2015)* investigated school segregation and found that American schools are highly segregated by race/ethnicity and socioeconomic status. They found that segregation is strongly associated with school behaviors and academic performance. The negative effects of segregation on school behavior and academic performance effect black and Hispanic students more as they are more likely to attend a segregated school. In fact, they found that black and Hispanic students were 3 and 2.5 times more likely than the average student to attend a highly segregated school (>90% minority enrollment). They noted in their study that segregated schools put a greater emphasis on discipline whereas schools serving middle class students put a greater emphasis on student initiative and creativity, which may contribute toward academic achievement.

Other research on the topic has concluded that reading for pleasure has a positive effect on academic performance as it can open up additional channels for knowledge that non-readers cannot access. *(Whitten, Labby, Sullivan, 2016)* examined the reading habits of sixty-five high school juniors, aged fifteen to seventeen years, at a rural Southeast Texas high school. While they only found marginal increase in English grades for pleasure readers versus non-readers they noticed significant increase in Maths and History in favour of pleasure readers.

(Hernandez, 2007) noted that many Hispanic students in the United States speak Spanish at home and also tend to perform worse academically when compared with other American students and investigated if there was a link between the two. The study found that home language use itself does not sufficiently explain academic achievement across different ethnic groups and pointed to other factors as being more significant. There have been many conflicting studies on this topic with some arguing that children who speak English at home have a marked advantage to perform better academically in an English speaking system. Other studies have noted that while many children from a Spanish speaking home tend to perform worse academically they put this lack of achievement down to socio economic factors as opposed to language.

I will conduct my study over a large ($N = 6504$) and nationally representative sample of students in the United States, which can be generalised for the United States. This study will look at the influence that the additional factors such as gender, ethnicity, language at home, relative age effect and reading for pleasure has over the students academic grades. I will then combine the most significant factors to produce multiple linear regression models and test if these models will be a significant predictor of academic grades.

Hypotheses of the Study

1. There will be no significant prediction of student $GradeScore$ by $Income$ and $MothersEducation$.

2. There will be no significant prediction of student $GradeScore$ by $Income$, $MothersEducation$, being an *Athlete*, $ReadingforPleasure$ and *Sex*

3. There will be no significant prediction of student $GradeScore$ by $Income$, $MothersEducation$, being an *Athlete*, $ReadingforPleasure$, *Sex* and *Ethnicity*.

Method

Participants

For this study I used Wave I of the `Add Health dataset` (N= 6504), which is available for public use and consists of survey responses from high school students in the United States http://www.cpc.unc.edu/projects/addhealth/design/wave1. Add Health is the largest and most comprehensive survey of adolescent health ever undertaken in the United States. The Wave I dataset consists of data collected from an In-Home interview conducted during the school year 1994-1995.

The Add Health dataset was compiled from survey data collected from 80 high schools across the United States selected to be nationally representative with respect to region of country, urbanicity, size, type, and ethnicity. Students in each school were then stratifed by grade and sex and chosen at random for an In-Home interview.

During the In-Home interview each student was asked questions from a questionnaire, which collected data from respondents across factors including social, economic, psychological and physical well being. A parent in each household, preferably the mother, was also surveyed answering questions covering topics such as education and employment, household income, neighbourhood, health and relationships.

Materials

GradeScore (n= 4899, M=307.3, SD= 59.8) is a numeric variable representing each students grades. It combines the grades from each of the 4 surveyed subjects: English, History, Science and Maths. A score of 100 was assigned for an A, 80 for a B, 60 for a C and 40 for a D in each subject. The maximum score a student can get indicating a perfect grade is 400. Only students who supplied grades for all 4 subjects are included in this study.

$MothersEducation$ (n= 5089) is an ordinal variable indicating level of education. For the purposes of building the multiple regression model this variable will be treated as numeric scale. It ranges from 0 (never went to school) through to 9 (professional training beyond a 4-year college or university) with various levels of education in between. A higher number indicates a higher level of education. The mode or most common level of education was 4 (n = 1464, high school graduate) with 7 (n = 1102, went to college, but did not graduate) the second most frequent.

$HouseholdIncome$ (n= 4929, M=47.7, SD= 56.3) is a numeric variable that ranges from 0 to 999 indicating the income of each household per year in thousands of US Dollars.

Athlete (n= 6498) is a categorical variable indicating a 1 (Athlete, n= 1582) or 0 (Non Athlete, n= 4916). The students were asked "*how many times they had played an active sport such as baseball, softball, basketball,* *soccer, swimming, or football in the past week?*". Students who answered "*five or more times*" were classified as an athlete.

ReadingFlag (*n*= 6497) is a categorical variable indicating a 1 (Reads for Pleasure, *n*= 1479) or 0 (Other, *n*= 5018). The students were asked "*During the past week, how many times did you do hobbies, such as playing a musical instrument reading, or doing arts and crafts?*". Students who answered "*five or more times*" were classified as being a reader for pleasure.

LanguageatHome (*n*= 6501) is a categorical variable indicating "English" (*n*= 6046) or "Non English" (*n*= 455) for students who answered to which language is spoken at home.

Ethnicity (*n*= 6477) is a categorical variable indicating the ethnic background of each student. The ethnic categories, of the population is divided into White (*n*= 3857), Black (*n*= 1567), Hispanic (*n*= 743), Asian (*n*= 224), Native American (*n*= 45) and Other (*n*= 41).

Sex (*n*= 6503) is a categorical variable indicating the gender of each student. The gender make up of the student population is "Female" (*n*= 3356) and "Male" (*n*= 3147).

Each of the numeric variables were standardised in preparation for the multiple regression model build and tests of skew and kurtosis were performed. Table 1 highlights the results. Skew and kurtosis for $GradeScore$ and $MothersEducation$ fell between 2 and -2 indicating normality. However, $Income$ was outside this range and so additional tests was required for this variable.

Table 1: Skew and Kurtosis for Numeric Variables

	Skew	Kurtosis
Grade Score	-0.312	-0.628
Income	8.776	119.725
Mothers Education	-0.272	-1.004

As an additional test, I tested the percentage of observations that fall outside ±1.96 and ±3.29 standard deviations from the mean. Results in Table 2 below indicate that less than 5% of observations fall outside this range meaning that I can treat this variable as normal for the purpose of this study.

Table 2: Additional Test of Normality

	No. Obs	No. ±1.96	% ±1.96	No. ±3.29	% ±3.29
Income	4929	93	0.019	50	0.01

After observing the distribution of $Income$, it became clear that a relatively small number of households were earning a very large income. I decided to remove these observations by standardising and capping all observations outside the ±3.29 range. The rationale behind this was so that extreme outliers would not skew the results and distort any trends when conducting the test.

Procedure

Hypothesis tests comparing the relationship between $Mother'sEducation$, $HouseholdIncome$ and the response variable *StudentGrades* have previously been carried out in Part I of the assignment. Table 3 shows the results of these correlation tests. The results suggests that Mothers with a higher level of education also produce children with significantly higher grades and the effect size indicates a small to medium effect. It also suggests that households with a higher level of annual income also produce children with significantly higher grades. The effect size indicates a small to medium effect.

Table 3: Correlation Tests

	Degree Freedom	r	p-value
Mothers Education	3875	0.229	<.001
Income	3713	0.24	<.001

A Students T test to compare the binary categorical variable *Athlete* versus *StudentGrades* was previously conducted in Part I of the assignment with the result of this in Table 4. New tests of significance for the additional variables versus the response variable *StudentGrades* was carried out as follows;

The first test will compare the relationship between the categorical variable *ReadingFlag* and *StudentGrades*. A Student's T test was used to test the difference between "Reader for Pleasure" and "Other" with respect to grades.

The second test will compare the relationship between the categorical variable *LanguageatHome* and *StudentGrades*. A Student's T test was used to test the difference between "English" and "Non English" with respect to grades.

The third test will compare the relationship between the categorical variable *Ethnicity* and *StudentGrades*. An ANOVA test was used to test the difference between the multi categorical ethnicity variable with respect to grades.

The fourth test will compare the relationship between the categorical variable *Sex* and *StudentGrades*. A Student's T test was used to test the difference between "Male" and "Female" with respect to grades.

Table 4: Student T Test

	Yes	No	t	p-value	d
Athlete - n	1305	3594	3.7	<.001	0.105
Athlete - M	312.6	305.5			
Athlete - SD	59.3	60.25			
Reads for Pleasure - n	1143	3756	-6.3	<.001	0.18
Reads for Pleasure - M	317.1	304.4			
Reads for Pleasure - SD	60.9	59.2			
English at Home - n	4601	298	0.73	0.46	0.021
English at Home - M	307.5	304.9			
English at Home - SD	59.9	59.2			
Female - n	2514	2385	11.96	<.001	0.343
Female - M	317.2	297			
Female - SD	57.6	60.48			

Results

The results of the T tests are show in Table 4. An *Athlete* engaging in regular team sports does have a small but positive effect on student grades. Students who read for pleasure also show a significant but higher effect size on student grades indicating students who read more can expect better grades. The test of the effect of English as the primary language spoken at home on student grades did not produce significant results indicating this has no bearing on academic achievement. The biggest effect size or contributor to academic achievement was in the test of gender versus student grades, which show a significant advantage of females over males with regard to academic achievement.

A one-way between subjects ANOVA was conducted to compare the effect of the six different ethnic categories on student grades. There was a significant effect on student grades at the p<.001 level across the different

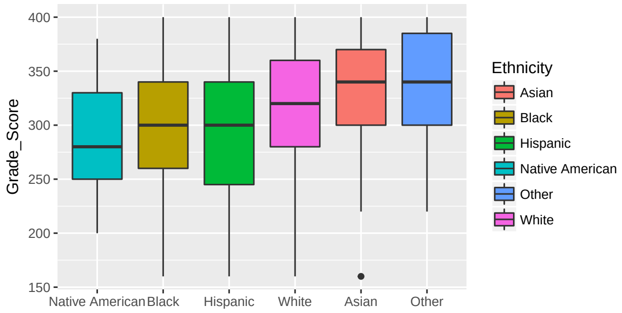

Figure 1: BoxPlot of Ethicity vs. Grade Score

ethnic groups [F(5, 4873) = 29.65, p = <.001]. Post hoc comparisons using the Tukey test indicated that the mean grade score for the Asian group (M = 329, SD = 56) was significantly different than for the White group (M = 314, SD = 61). Both these groups were also significantly different from the Black (M = 295, SD = 56), Hispanic (M = 294, SD = 57) and Native American (M = 278, SD = 63). However the Black, Hispanic and Native American groups were not significantly different from each other. Also, the Other group (M = 321, SD = 68) was not significantly different from any other group suggesting the observed differences here are due to random chance. Taken together, these results suggest that ethnicity really does have an effect on academic achievement. Specifically, Asians are the highest academic achievers, followed by White with Black, Hispanic and Native American all following with regard to academic achievement.

Model 1

A multiple linear regression was calculated to predict a student's academic *GradeScore* based on the continuous variables $MothersEducation$ and $HouseholdIncome$. A significant regression equation was found (F(2,2850)= 116.4, p= <.001) with an adjusted R^2 of .075. Students predicted *GradeScore* is equal to 0.054 + 0.154($MothersEducation$) + 0.305($HouseholdIncome$), where both $MothersEducation$ and $HouseholdIncome$ are both standardised continuous variables. The students standardised *GradeScore* increases 0.305 standard deviations for each standard deviation increase in $HouseholdIncome$ and increases 0.154 standard deviations for each standard deviation increase in $MothersEducation$. Both $HouseholdIncome$ (p< .001) and $MothersEducation$ (p< .001) were significant predictors of *GradeScore*.

An analysis of Cooks distance was carried out, which showed that the data contained no extreme outliers and none with Cook's distance >1 as outlined in *(Field (2013))* (Max Cooks Distance = 0.02). Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern ($HouseholdIncome$, VIF= 1.20; $MothersEducation$, VIF= 1.20) and were within acceptable levels (VIF <2.5) as outlined in *(Tarling (2008))*. The density plot of standardised residuals indicated that the data contained normally distributed errors, as did the normal Q-Q plot of standardised residuals, which showed all points were close to the line and plotted in Figure 2. The data also meets the assumption of non-zero variances of the predictors.

Model 2

A multiple linear regression was calculated to predict a student's academic *GradeScore* based on the continuous variables $MothersEducation$ and $HouseholdIncome$ and nominal variables $ReadsforPleasure$, *Sex* and *Athlete*. The nominal variables were recoded as follows; $ReadsforPleasure$ (1= reads, 0= other), *Sex*(1= Female, 0= Male), *Athlete* (1=Athlete, 0= Non Athlete).

A significant regression equation was found (F(5,2847)= 74.19, p= <.001) with an adjusted R^2 of .113. Students predicted *GradeScore* is equal to -0.223 + 0.149($MothersEducation$) + 0.287($HouseholdIncome$) + 0.165($ReadsforPleasure$) + 0.377(*Sex*) + 0.144(*Athlete*). The students standardised *GradeScore*

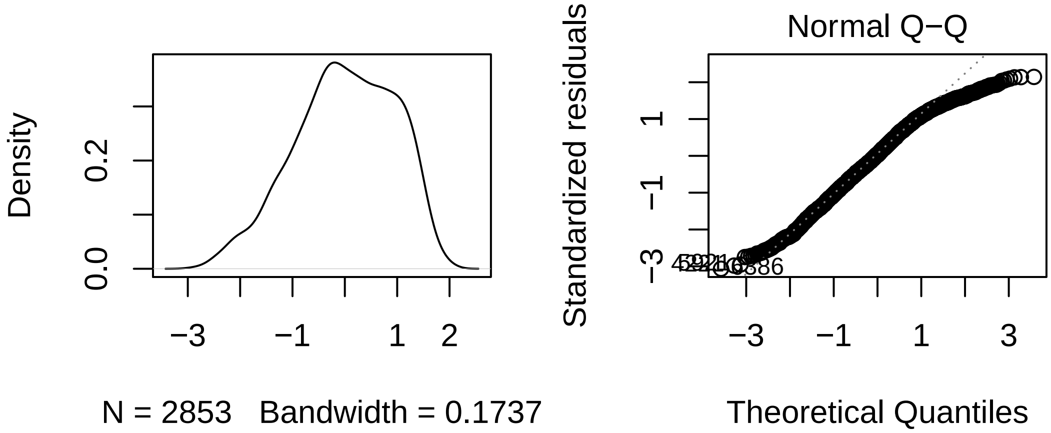

Figure 2: Model 1: Density Plot of model residuals (left) QQ Plot of model residuals (right)

increases 0.287 standard deviations for each standard deviation increase in $HouseholdIncome$ and increases 0.149 standard deviations for each standard deviation increase in $MothersEducation$. Students who $ReadsforPleasure$ have a $GradeScore$ 0.165 standard deviations higher, Females have a $GradeScore$ 0.377 standard deviations higher and Athletes have a $GradeScore$ 0.144 standard deviations higher. All variables were significant predictors of $GradeScore$.

An analysis of Cooks distance was carried out, which showed that the data contained no extreme outliers and none with Cook's distance >1 as outlined in *(Field (2013))* (Max Cooks Distance = 0.013). Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern ($HouseholdIncome$, VIF= 1.20; $MothersEducation$, VIF= 1.20; $ReadsforPleasure$, VIF= 1.04; Sex, VIF = 1.05; $Athlete$, VIF= 1.08) and were within acceptable levels (VIF <2.5) as outlined in *(Tarling (2008))*. The density plot of standardised residuals indicated that the data contained normally distributed errors, as did the normal Q-Q plot of standardised residuals, which showed all points were close to the line. The data also meets the assumption of non-zero variances of the predictors.

Model 3

A multiple linear regression was calculated to predict a student's academic $GradeScore$ based on the continuous variables $MothersEducation$ and $HouseholdIncome$ and nominal variables $ReadsforPleasure$, Sex, $Athlete$ and $Ethnicity$. The nominal variables were recoded as follows; $ReadsforPleasure$ (1= reads, 0= other), Sex(1= Female, 0= Male), $Athlete$ (1=Athlete, 0= Non Athlete). In order to capture $Ethnicity$ it was recoded into 2 variables; Asian_dummy (1= Asian, 0= all other) and White_dummy (1=White, 0= all other).

A significant regression equation was found (F(7,2815)= 59.01, p= <.001) with an adjusted R^2 of .126. Students predicted $GradeScore$ is equal to -0.38366 + 0.145($MothersEducation$) + 0.244($HouseholdIncome$) + 0.153($ReadsforPleasure$) + 0.382(Sex) + 0.158($Athlete$) + 0.369($Asian$) + 0.222($White$). The students standardised $GradeScore$ increases 0.244 standard deviations for each standard deviation increase in $HouseholdIncome$ and increases 0.145 standard deviations for each standard deviation increase in $MothersEducation$. Students who $ReadsforPleasure$ have a $GradeScore$ 0.153 standard deviations higher, Females have a $GradeScore$ 0.382 standard deviations higher and Athletes have a $GradeScore$ 0.158 standard deviations higher, students of $Asian$ and White ethnicity can expect to score 0.369 and 0.222 standard deviations higher when compared to black, hispanic and native american ethnicities. All variables were significant predictors of $GradeScore$.

An analysis of Cooks distance was carried out, which showed that the data contained no extreme outliers and none with Cook's distance >1 as outlined in *(Field (2013))* (Max Cooks Distance = 0.015). Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern ($HouseholdIncome$, VIF= 1.25; $MothersEducation$, VIF= 1.20; *ReadsforPleasure*, VIF= 1.04; Sex, VIF = 1.05; *Athlete*, VIF= 1.08; $Asian$, VIF = 1.06; $White$, VIF = 1.12) and were within acceptable levels (VIF <2.5) as outlined in *(Tarling (2008))*. The density plot of standardised residuals indicated that the data contained normally distributed errors, as did the normal Q-Q plot of standardised residuals, which showed all points were close to the line. The data also meets the assumption of non-zero variances of the predictors.

Test Data

The original data set was partitioned on 85% to a training dataset (n= 5528) and 15% to a test datset (n= 976). The models were built on the train dataset. As an additional validation I ran Model 3 against the test dataset to produce a set of predicted *GradeScores* for each student. A correlation test of predicted vs. actual *GradeScores* yielded a significant positive association ($r(512) = .387$, $p < .001$).

Conclusion and Discussion

The aim of this paper was to build multiple linear regression models capable of predicting student grades better than just taking the mean. The initial variable exploration indicated results consistent with previous studies. The test of difference between gender showed females with significantly higher grades versus male backing up research by *(Sayid Dabbagh Ghazvinia, Milad Khajehpour, 2011)*. The ANOVA test across different ethnic groups also concurred with *(Palardy, Rumberger, Butler, 2015)* findings in that Black and Hispanic students tend to perform worse academically when compared with White and Asian students. *(Hernandez, 2007)* noted that there had been many conflicting studies on the effect of living in a non-English speaking home and academic performance. The relationship test between language spoken at home and academic grades was not significant indicating no relationship between these for the tested population. *(Whitten, Labby, Sullivan, 2016)* found that students who read for pleasure tend to perform better academically and the test conducted in this study was significant with this finding.

Once the significant variables were identified a number of models were built to predict academic achievement. Model 1 was constructed using only $MothersEducation$ and $HouseholdIncome$ as predictor variables with increases in $HouseholdIncome$ having twice as large an effect on academic grades than a similar increase in $MothersEducation$. This model was significant in that it could predict better than just taking the average and was capable of explaining 7.5% of the variance in student grades.

Model 2 built upon the base of Model 1, adding additional significant variables such as $ReadsforPleasure$, Sex and *Athlete* to the model. The addition of these variables raised the variability that could be explained by the model to 11.3%. The model could show that Female students can expect to score in the region of 0.377 standard deviations higher than males in academic tests. A similar but smaller effect was shown for students who read for pleasure and those who regularly engage in team sports.

The final model built was Model 3 which added *Ethnicity* to the equation. The ANOVA test indicated 3 significant groups of academic achievement: Asian , White and a minority group including Black & Hispanic students. Recoding dummy variables that reflected that 3 categories of ethnicity found that Asian students can expect to score 0.369 standard deviations higher than the minority group in academic tests while White students can expect a similar but smaller effect. The inclusion of the additional variables increased the variability that could be explained by the model to 12.6%.

The 3 hypothesis tests set out at the start of the paper stated that the models would not be predictive with regard to *StudentGrades* and that the null hypothesis would be true. The results show that the null hypothesis is false and that the models are both significant and predictive. Running Model 3 against an unknown test dataset also yielded positive results with the variability that could be explained by the model rising to 14.4% for the test dataset.

However, as the model can only explain about 14% of the variability, the model is not accurate enough to be applied in a practical setting to actually predict student grades. However, some of the insights discovered in the creation of this model can offer value. For example, an Asian female from a high income family can expect significantly better grades than average. While conversely, a Black male from a low income family can expect significantly lower grades than average. This level of insight can help inform boards of education and government policy makers as to who the “at risk” groups are and where to divert funding and resources in an effort to raise educational standards.

References

Sayid Dabbagh Ghazvinia , Milad Khajehpou ,“Gender differences in factors affecting academic performance of high school students” (2011)

Palardy, Rumberger, Butler, “The effect of high school socioeconomic, racial, and linguistic segregation on academic performance and school behaviors” (2015)

Whitten, Labby, Sullivan, “The impact of Pleasure Reading on Academic Success” (2016)

Hernandez, “Home Language Use and Hispanic Academic Achievement: Evidence from Texas High Schools” (2007)

Field, “Discovering Statistics using IBM SPSS Statistics” (2013)

Tarling, “Statistical Modelling for Social Researchers” (2008)

Appendix

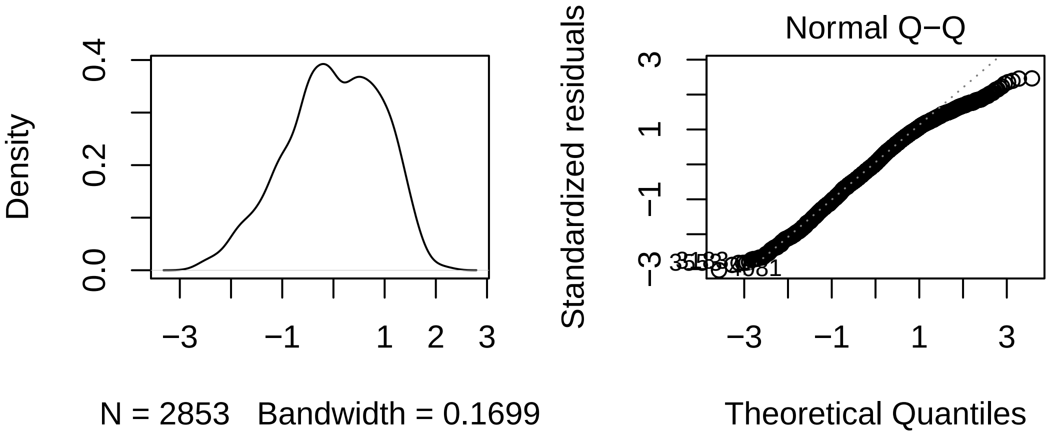

Figure 3: Model 2: Density Plot of model residuals (left) QQ Plot of model residuals (right)

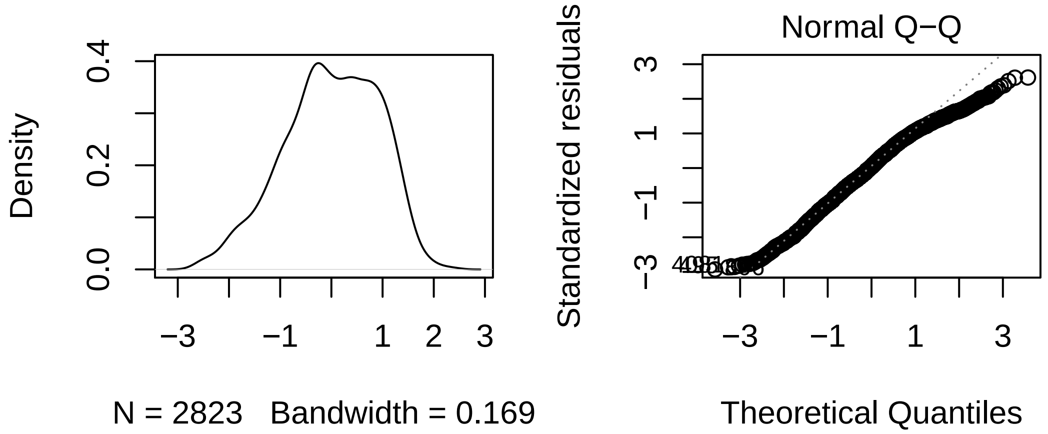

Figure 4: Model 3: Density Plot of model residuals (left) QQ Plot of model residuals (right)

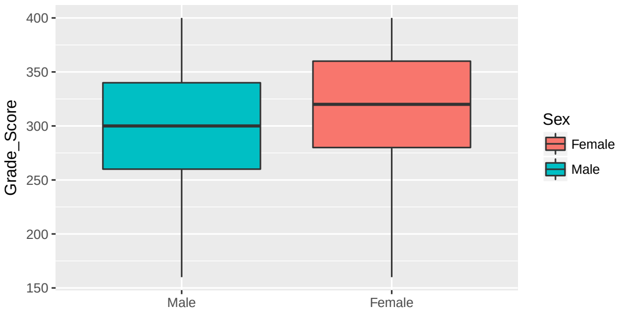

Figure 5: BoxPlot of Sex vs. Grade Score