Final Project : Student Performance in Exams

MATH60604A - Statistical Modelling

Submitted to : Prof. Juliana Schulz

Presented by:

Zihang Cai - 11319479

Helen Ma - 11313446

Jiahua Shang - 11319456

Danni Weng - 11304053

Introduction

Consider the “exams.csv” data (from https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams (https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams)). The dataset includes test scores from 1000 kids in reading, writing, and math, as well as some additional data about the students. The following variables are included in the file:

Variable	Description
gender	Female or Male
race.ethnicity	Race/Ethnicity group A, B, C, D, E
parental.level.of.education	Parents’ education level: some high school, some college, high school, associate’s degree, bachelor’s degree, master’s degree
lunch	Student’s meal plan: reduced/free or standard
test.preparation.course	Indicator if the student took the preparation course prior to the test (completed or none)
math.score	Test score in math
reading.score	Test score in reading
writing.score	Test score in writing

```
##     gender    race.ethnicity     parental.level.of.education
##  female:483   group A: 79    associate's degree:203
##  male  :517   group B:205    bachelor's degree :112
##               group C:323    high school       :202
##               group D:262    master's degree   : 70
##               group E:131    some college      :222
##                              some high school  :191
##           lunch     test.preparation.course   math.score    reading.score
##  free/reduced:348   completed:335           Min.   : 13.0   Min.   : 27
##  standard    :652   none     :665           1st Qu.: 56.0   1st Qu.: 60
##                                             Median : 66.5   Median : 70
##                                             Mean   : 66.4   Mean   : 69
##                                             3rd Qu.: 77.0   3rd Qu.: 79
##                                             Max.   :100.0   Max.   :100
##  writing.score
##  Min.   : 23.00
##  1st Qu.: 58.00
##  Median : 68.00
##  Mean   : 67.74
##  3rd Qu.: 79.00
##  Max.   :100.00
```

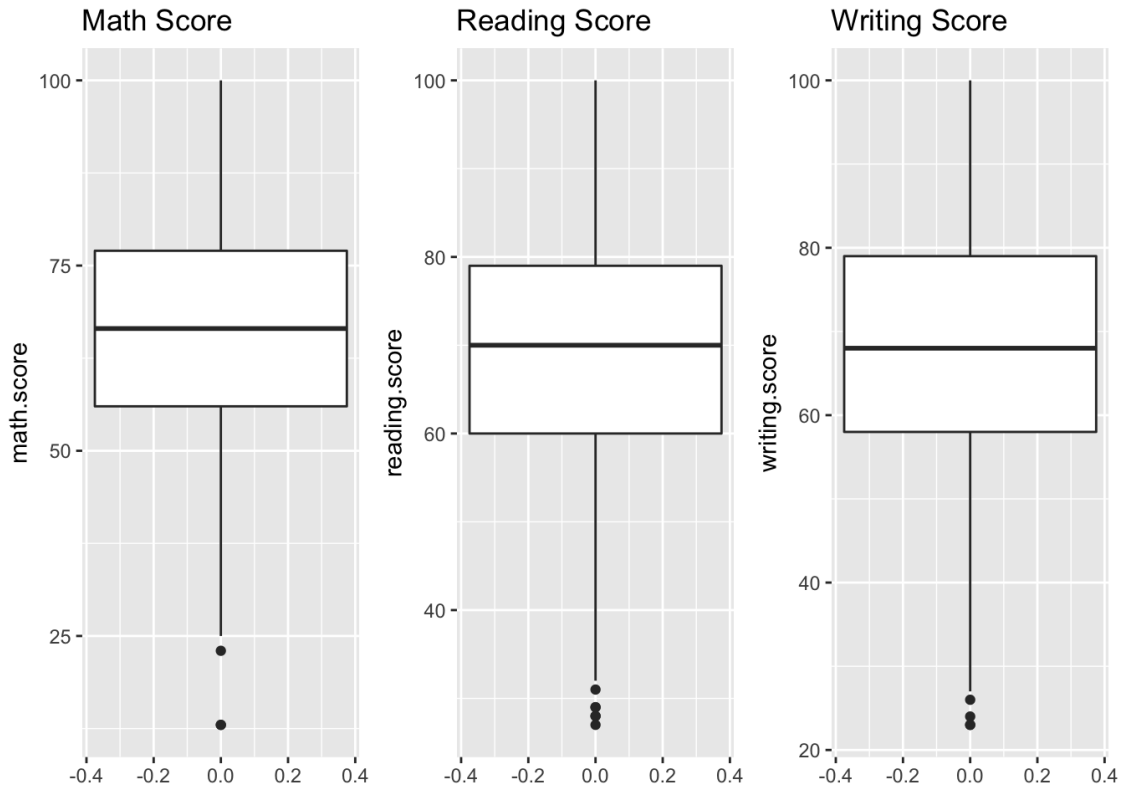
Data Exploration

From the summary above, the dataset is fairly structured and does not contain any NULL rows at a first glance. We will now explore each variable more in details.

Score

From the summary, we can see that the data contains three numeric variables of scores for each subject : `math.score` , `reading.score` and `writing.score` . All variables are capped at a maximum score of 100% and this seem coherent. The mean score in math, reading and writing are 66.4%, 69% and 67.74% respectively and for the purpose of this analysis, we will be creating a new variable `avg_score` , which is the average of the three score of three scores as the response variable.

```
##    math.score    reading.score writing.score
##  Min.   : 13.0   Min.   : 27   Min.   : 23.00
##  1st Qu.: 56.0   1st Qu.: 60   1st Qu.: 58.00
##  Median : 66.5   Median : 70   Median : 68.00
##  Mean   : 66.4   Mean   : 69   Mean   : 67.74
##  3rd Qu.: 77.0   3rd Qu.: 79   3rd Qu.: 79.00
##  Max.   :100.0   Max.   :100   Max.   :100.00
```

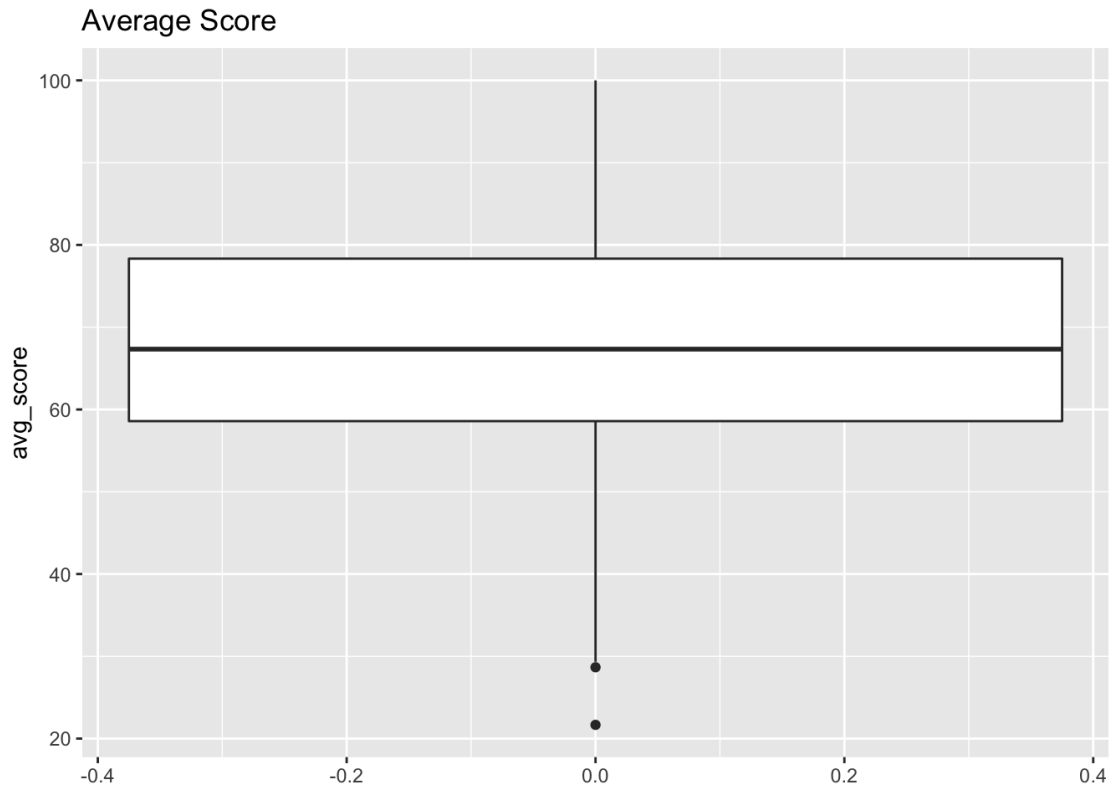


The code to create the new variable `avg_score` is shown below:

```
data<-raw_data[,1:5]
#create new variable Avg Score
data$avg_score<-(raw_data$math.score+raw_data$reading.score+raw_data$writing.score)/3
```

The variable `avg_score` ranges from a minimum score of 21.67% to maximum score of 100%. From boxplot, the data points seem to be distributed nicely in this range as the median is fairly close to the mean value. Moreover, the upper quartile at 58.6% and lower quartile at 78.3% seem to be roughly symmetric distance from the median, which seems like a plausible distribution for students tests scores. There are a few points that may be further away from the center (scores of 20%-30%), but they do not seem too problematic.

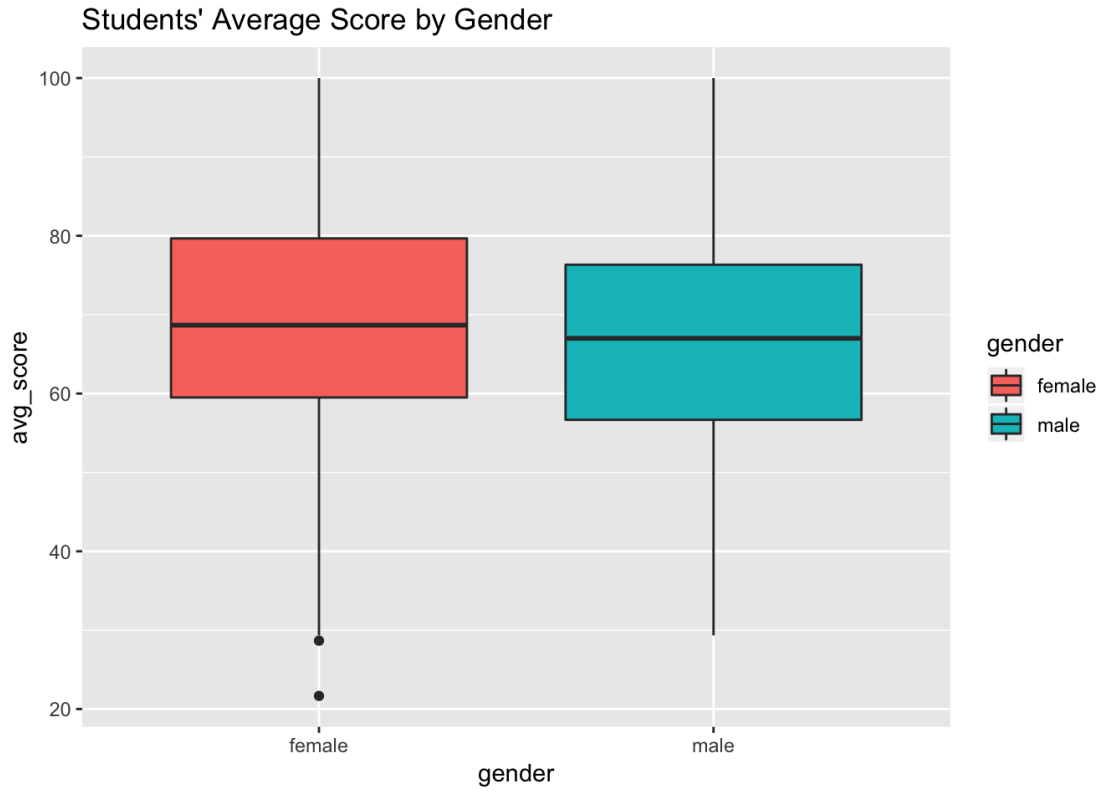
```
##     gender    race.ethnicity     parental.level.of.education
##  female:483   group A: 79    associate's degree:203
##  male  :517   group B:205    bachelor's degree :112
##               group C:323    high school       :202
##               group D:262    master's degree   : 70
##               group E:131    some college      :222
##                              some high school  :191
##           lunch     test.preparation.course   avg_score
##  free/reduced:348   completed:335           Min.   : 21.67
##  standard    :652   none     :665           1st Qu.: 58.58
##                                             Median : 67.33
##                                             Mean   : 67.71
##                                             3rd Qu.: 78.33
##                                             Max.   :100.00
```



Gender

The variable `Gender` is a categorical variable with two levels : Female or Male. There are 483 observations of female students and 517 observations of male students in the dataset. This distribution seem to be balanced between both genders. When analyzing the average scores by gender, we can observe that the score distribution seem slightly higher for female students, but the differences are slight. Moreover, the lowest scores noticed earlier seem to belong to female students.

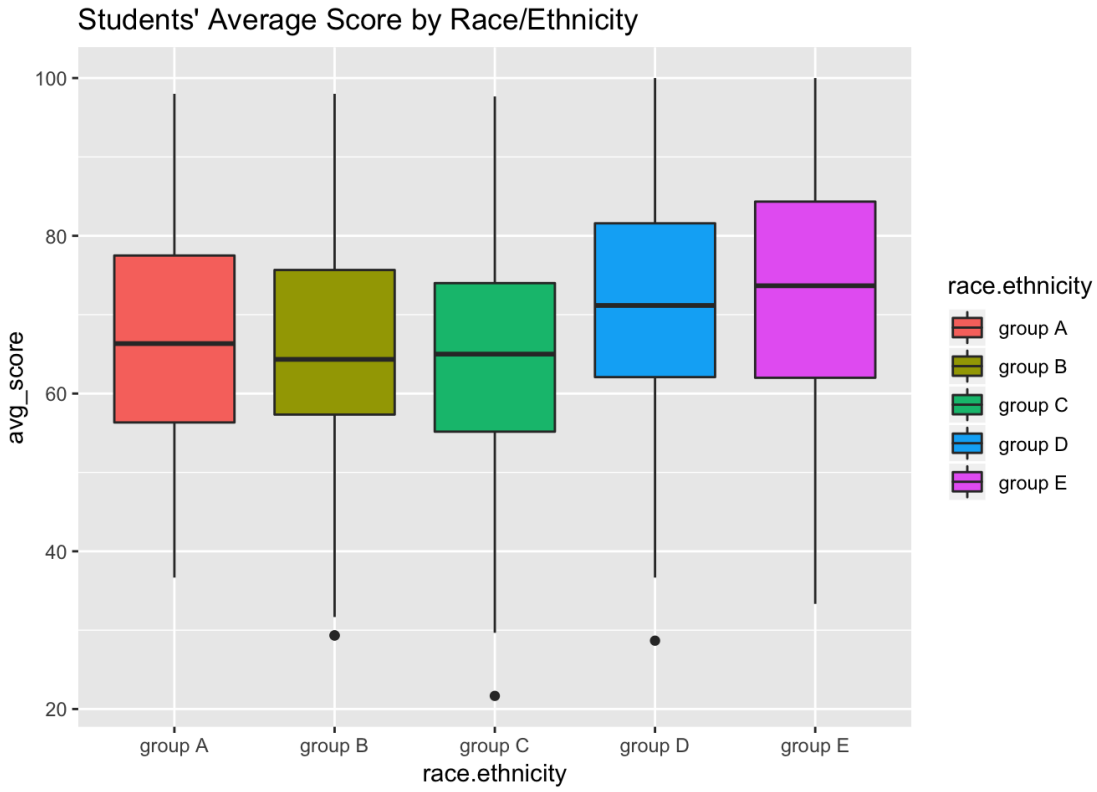
```
##        frequencies percentage cumulativepercentage
## female         483       48.3                 48.3
## male           517       51.7                100.0
## Totals        1000      100.0                100.0
```



Race/Ethnicity

The variable `Race/Ethnicity` is a categorical variable grouped into five levels : group A, B, C, D and E. From the frequency table, we can observe that there's an uneven distribution of observations among ethnicity group: Group C represents the largest proportion at 32.3% and group C only represents 7.9% of total observations. Given that we know the actual distribution of students population data, it could be possible that this dataset has an under-representation of Group A and over-representation of group C, which we have to keep in mind when drawing any conclusions or interpretations later.

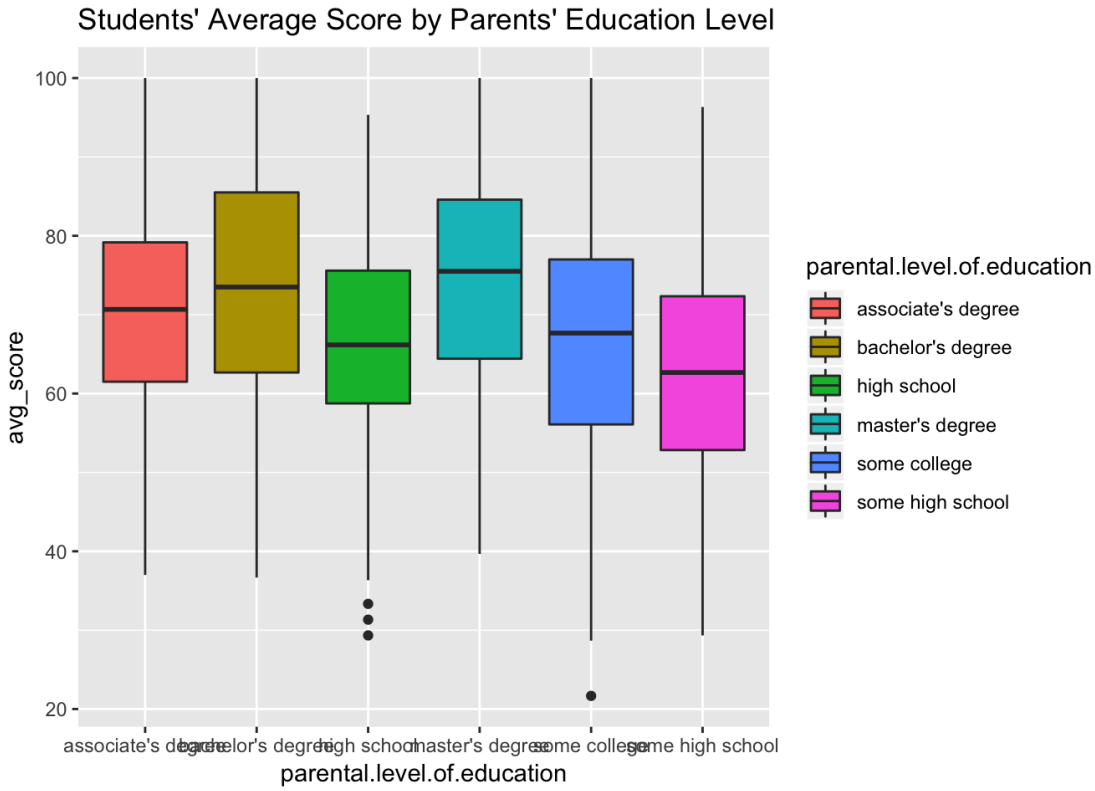
```
##         frequencies percentage cumulativepercentage
## group A          79        7.9                  7.9
## group B         205       20.5                 28.4
## group C         323       32.3                 60.7
## group D         262       26.2                 86.9
## group E         131       13.1                100.0
## Totals         1000      100.0                100.0
```



Parent's education level

The variable `Parental.level.of.education` is a categorical variable grouped into six levels : some high school, high school, some college, associate's degree, bachelor's degree and master's degree. From the frequency table, we can observe that the distributions of observations are roughly similar for each group, except for the higher levels of education (bachelor and master's degree). Given that this variable has somewhat a ordinal structure (for example, a high school diploma is lower education level than a college degree, which is lower than a master degree), we decide to group adjacent levels of educations levels together and condensed them into three levels.

```
##                   frequencies percentage cumulativepercentage
## associate's degree         203       20.3                 20.3
## bachelor's degree          112       11.2                 31.5
## high school                202       20.2                 51.7
## master's degree             70        7.0                 58.7
## some college               222       22.2                 80.9
## some high school           191       19.1                100.0
## Totals                    1000      100.0                100.0
```



Therefore, a new variable `parent_educ` will be created with three levels:

1: high school and some high school

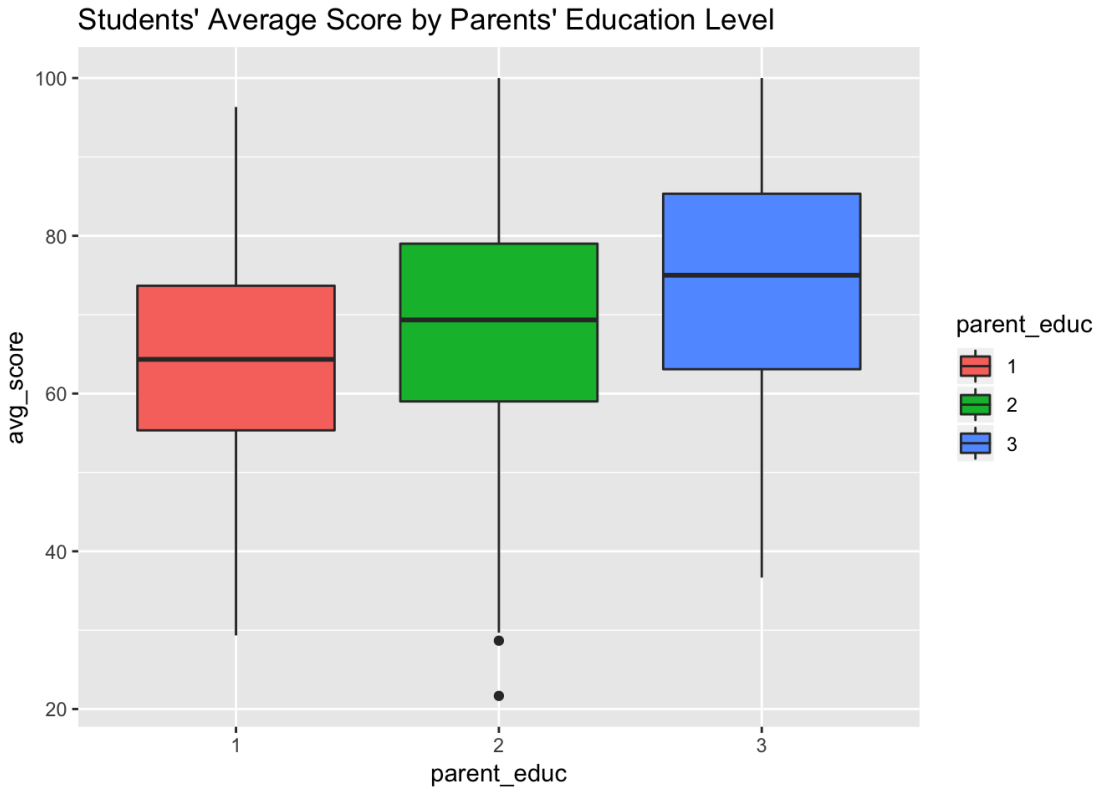
2: some college and associate's degree

3: bachelor's and master's degree

```
data$parent_educ <- data$parental.level.of.education
levels(data$parent_educ) <- list("1" = c("high school", "some high school"),
                                 "2" = c("some college", "associate's degree"),
                                 "3" = c("bachelor's degree", "master's degree"))
data$parent_educ<-relevel(as.factor(data$parent_educ),ref="1")
```

The new categories are distributed as follows:

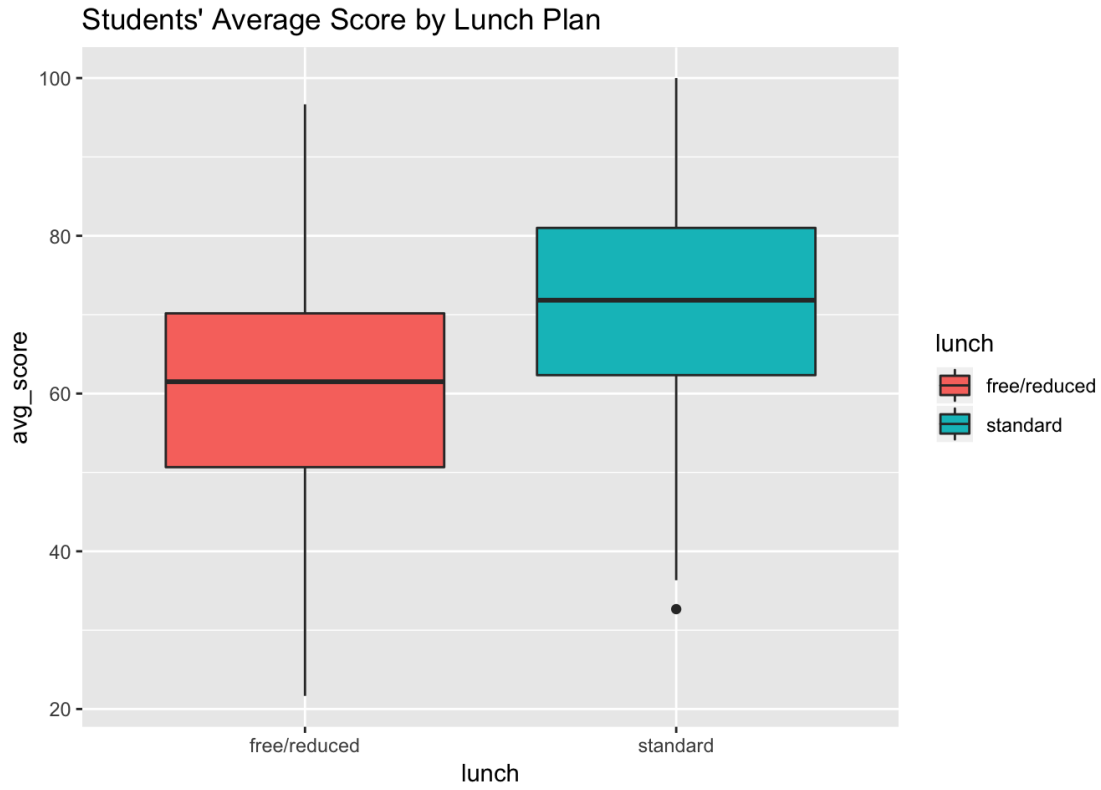
```
##         frequencies percentage cumulativepercentage
## 1               393       39.3                 39.3
## 2               425       42.5                 81.8
## 3               182       18.2                100.0
## Totals         1000      100.0                100.0
```



Lunch Meal Plan

The variable `Lunch` is a categorical variable with two levels describing the option of lunch plan the student has selected at school. The first option is the free or reduced meal plan and the other alternative is the standard meal plan. Looking at the distribution of observations, it consists of 34.8% of the “free/reduced” plan and 65.2% observations of standard lunch plan. Again, the standard plan seem to be slightly over-represented in this data set. When looking at the plot of average scores by lunch plan selected, it seems to suggest that students who selected the standard lunch plan tend to score higher than students who selected the reduced plan, but additional investigation is required to assess whether this difference is statistically significant.

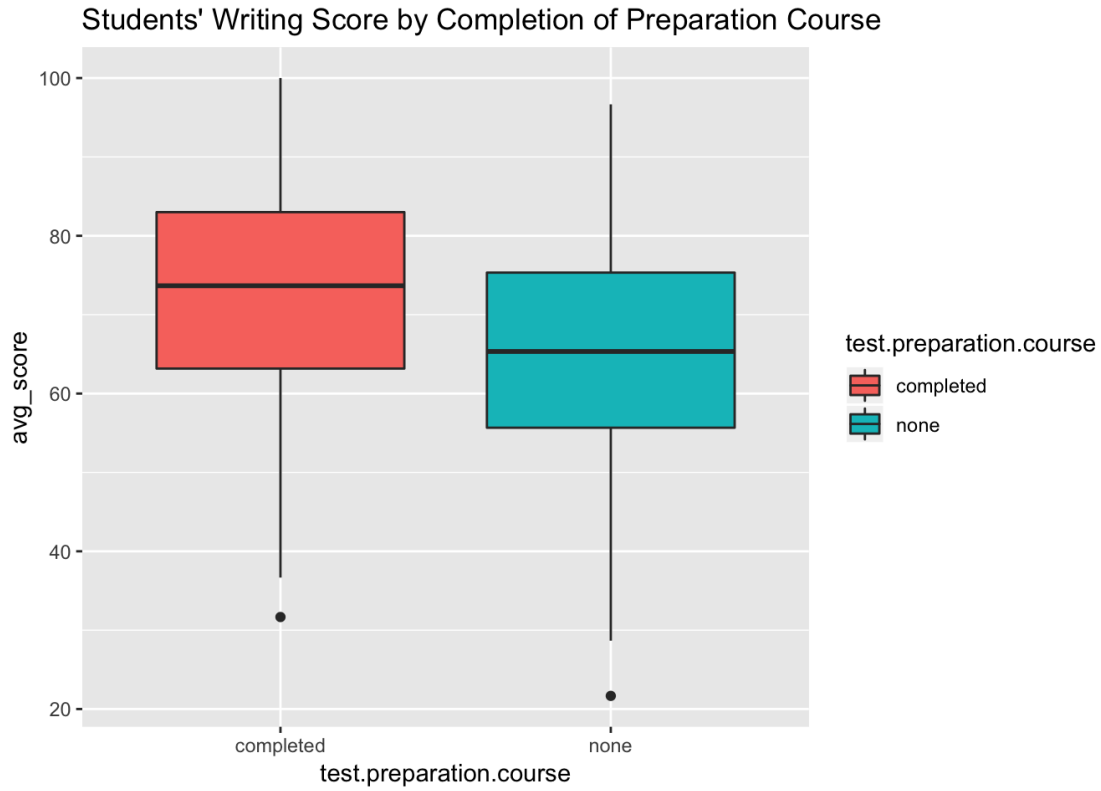
```
##              frequencies percentage cumulativepercentage
## free/reduced         348       34.8                 34.8
## standard             652       65.2                100.0
## Totals              1000      100.0                100.0
```



Test preparation course

The last variable is an indicator variable flagging whether or not the student has completed the test preparation course prior to the exam. Looking at the distribution of data, it seems that only 33.5% of students in the dataset have completed the preparation course. We will recode this variable to take value 1 if the student has completed the preparation course and value 0 otherwise.

```
##           frequencies percentage cumulativepercentage
## completed         335       33.5                 33.5
## none              665       66.5                100.0
## Totals           1000      100.0                100.0
```



```
#Changing indicator variable for test preparation course, if completed = 1, otherwise 0
data$prep_course<-as.factor(ifelse(data$test.preparation.course=="completed", 1,0))
```

Final data

Following all the modification to the raw data above, the final dataset contains the following variables.

```
mydata<-data[,c(6,1,2,7,4,8)]
summary(mydata)
```

```
##    avg_score         gender    race.ethnicity parent_educ
##  Min.   : 21.67   female:483   group A: 79    1:393
##  1st Qu.: 58.58   male  :517   group B:205    2:425
##  Median : 67.33                group C:323    3:182
##  Mean   : 67.71                group D:262
##  3rd Qu.: 78.33                group E:131
##  Max.   :100.00
##          lunch     prep_course
##  free/reduced:348   0:665
##  standard    :652   1:335
##
##
##
##
```

```
#write.csv(mydata,"Data/exam_final.csv")
```

Variable	Description
avg_score	Average score in math, reading and writing
gender	Female or Male
race.ethnicity	Race/Ethnicity group A, B, C, D, E
parent_educ	Parents' education level: 1: some high school/high school, 2:some college/associate's degree, 3:bachelor's degree/master's degree
lunch	Student's meal plan: reduced/free or standard
prep course	Indicator if the student took the preparation course prior to the test (1 if completed, 0 otherwise)

Question 1

We would like to investigate the factors influencing a student's score. Beginning by fitting a linear regression with all variables to the response variable `avg_score` .

a. Provide the fitted model (group A as reference level of race.ethnicity; some high school as reference level of parent_educ), and check whether 'avg_score' is well explained by this model.

```
## Make categorical variables and relevel
mydata1<-mydata
mydata1$race.ethnicity<-relevel(as.factor(mydata1$race.ethnicity),ref=1)
mydata1$parent_educ<-relevel(as.factor(mydata1$parent_educ),ref=1)
mydata1$lunch<-as.factor(mydata1$lunch)
mydata1$prep_course<-as.factor(mydata1$prep_course)

## Linear regression on all variables
mod1a<-lm(avg_score~.,data=mydata1)
summary(mod1a)
```

```
##
## Call:
## lm(formula = avg_score ~ ., data = mydata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.339  -8.858   0.195   8.882  27.565
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              53.9983     1.6641  32.448  < 2e-16 ***
## gendermale               -2.4922     0.7856  -3.172  0.00156 **
## race.ethnicitygroup B    -0.6441     1.6327  -0.394  0.69330
## race.ethnicitygroup C    -0.5632     1.5488  -0.364  0.71623
## race.ethnicitygroup D     5.4746     1.5838   3.457  0.00057 ***
## race.ethnicitygroup E     8.3747     1.7595   4.760 2.23e-06 ***
## parent_educ2              4.5701     0.8660   5.277 1.61e-07 ***
## parent_educ3              9.3658     1.1062   8.467  < 2e-16 ***
## lunchstandard            10.2843     0.8217  12.516  < 2e-16 ***
## prep_course1              7.2610     0.8269   8.781  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.29 on 990 degrees of freedom
## Multiple R-squared:  0.2891, Adjusted R-squared:  0.2827
## F-statistic: 44.74 on 9 and 990 DF,  p-value: < 2.2e-16
```

According to the output, the fitted model can be expressed as:

$$\hat{avgscore} = 53.9983 - 2.4922gender_{male} - 0.6441race.ethnicity_{groupB} - 0.5632race.ethnicity_{groupC} + 5.4746race.ethnicity_{groupD} + 8.3747ace.ethnicity_{groupE} + 4.5701parenteduc_2 + 9.3658parenteduc_3 + 10.2843lunch_{standard} + 7.2610prepcourse$$

The value of R^2 is 0.2891 and adjusted R^2 is 0.2827, which indicates that this model explains only 28-29% of the total variation. Since $R^2 = (SST - SSE)/SST = SSR/SST$ and $SST = SSE + SSR$, it shows that the $SSR < SSE$ and suggests that this model poorly explain the variation in the response variable avg_score.

b. Is race.ethnicity globally significant? Compare group C to group A and interpret the differences found?

```
library(car)
## To do a global effect test
Anova(mod1a,type=3)
```

```
## Anova Table (Type III tests)
##
## Response: avg_score
##                Sum Sq  Df  F value    Pr(>F)
## (Intercept)    159110   1 1052.892 < 2.2e-16 ***
## gender           1521   1   10.064  0.001558 **
## race.ethnicity  12228   4   20.230 4.943e-16 ***
## parent_educ     11471   2   37.953 < 2.2e-16 ***
## lunch           23674   1  156.661 < 2.2e-16 ***
## prep_course     11651   1   77.101 < 2.2e-16 ***
## Residuals      149606 990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For variable race.ethnicity, to carry out a global test, we set:

$H_0 : \beta_{race.ethnicitygroupB} = \beta_{race.ethnicitygroupC} = \beta_{race.ethnicitygroupD} = \beta_{race.ethnicitygroupE} = 0$

H_1: at least one of $\beta_{race.ethnicitygroupB}$ or $\beta_{race.ethnicitygroupC}$ or $\beta_{race.ethnicitygroupD}$ or $\beta_{race.ethnicitygroupE} \neq 0$

According to the output, we can find the F-value (20.230) and p-value ($4.943 * 10^{-16}$) of race.ethnicity is smaller than any reasonable α , then we can reject H_0. Therefore, variable race.ethnicity is globally significant in this model.

According the summary table given in part a), the estimated coefficient $\hat{\beta}_{race.ethnicitygroupC} = -0.5632$. It indicates that, on average, the differences in average score of race/ethnicity group A and group C is -0.5632 marks, holding all other variables constant.

c. Fit a new linear model including an interaction between `gender` ***and*** `lunch` ***and all other variables as well and provide the fitted model. Then justify whether the lunch type influences the effect of gender on average score significantly.***

```
## Fit a new model
mod1c<-lm(avg_score~gender*lunch+race.ethnicity+parent_educ+prep_course,data=mydata1)
summary(mod1c)
```

```
##
## Call:
## lm(formula = avg_score ~ gender * lunch + race.ethnicity + parent_educ +
##     prep_course, data = mydata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.814  -8.736   0.191   8.875  27.111
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                54.4737     1.7637  30.886  < 2e-16 ***
## gendermale                 -3.3697     1.3332  -2.528  0.01164 *
## lunchstandard               9.5815     1.1914   8.042 2.51e-15 ***
## race.ethnicitygroup B      -0.6848     1.6337  -0.419  0.67520
## race.ethnicitygroup C      -0.5846     1.5493  -0.377  0.70600
## race.ethnicitygroup D       5.4350     1.5848   3.429  0.00063 ***
## race.ethnicitygroup E       8.3986     1.7601   4.772 2.10e-06 ***
## parent_educ2                4.5919     0.8666   5.299 1.44e-07 ***
## parent_educ3                9.4021     1.1073   8.491  < 2e-16 ***
## prep_course1                7.2704     0.8271   8.790  < 2e-16 ***
## gendermale:lunchstandard    1.3374     1.6415   0.815  0.41540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.3 on 989 degrees of freedom
## Multiple R-squared:  0.2896, Adjusted R-squared:  0.2824
## F-statistic: 40.32 on 10 and 989 DF,  p-value: < 2.2e-16
```

According to the output, the fitted model can be expressed as:

$$\hat{avgscore} = 54.4737 - 3.3697gender_{male} - 0.6848race.ethnicity_{groupB} - 0.5846race.ethnicity_{groupC} + 5.4350race.ethnicity_{groupD} + 8.3986race.ethnicity_{groupE} + 4.5919parenteduc_2 + 9.4021parenteduc_3 + 9.5815lunch_{standard} + 7.2704prepcourse + gender_{male} * lunch_{standard}$$

To find whether lunch type influences the effect of gender on score, we can check if interaction is significant, since these two variables are binary, a hypothesis test can be set as:

$H_0: \beta_{gendermale*lunchstandard} = 0$

$H_1: \beta_{gendermale*lunchstandard} \neq 0$

From the output, the result of the test statistic is T=0.815 with a corresponding p-value of 0.41540 which is greater than any reasonable α thus we cannot reject null hypothesis; therefore we cannot conclude that lunch influences the effect of gender significantly.

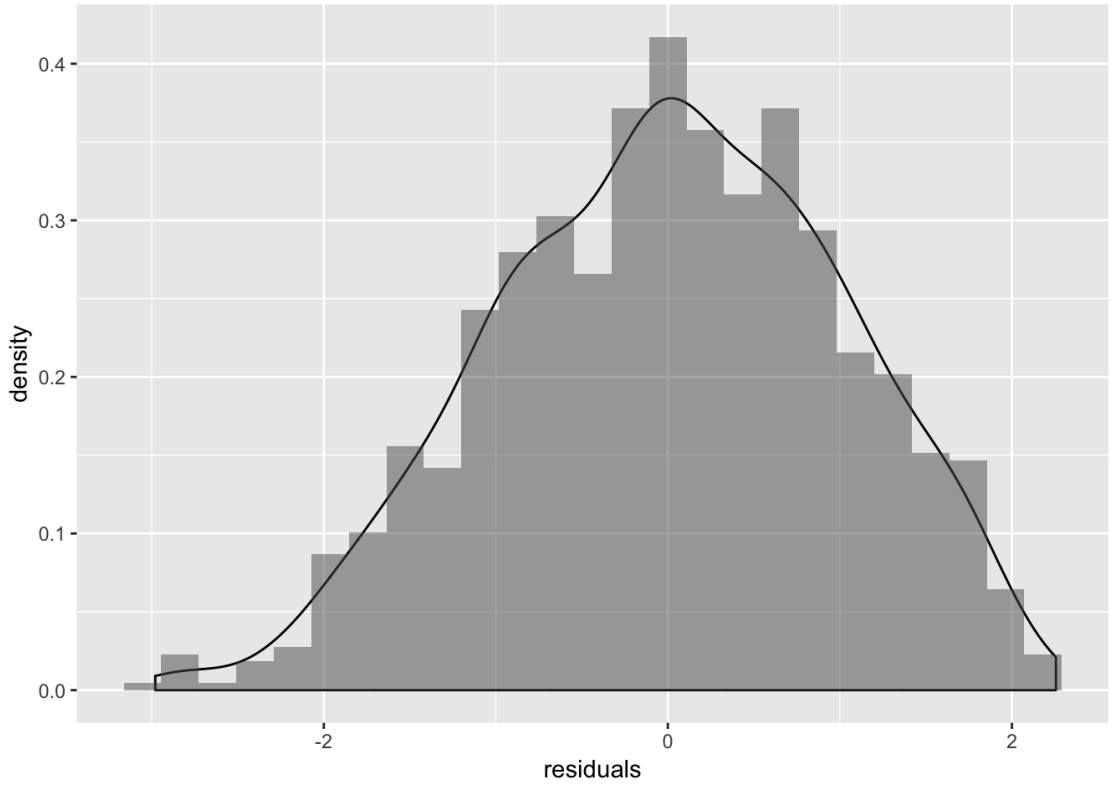
d. Formally test if the interpretations to above questions are valid. Carry a residual analysis of the model in part a) and comment on the results.

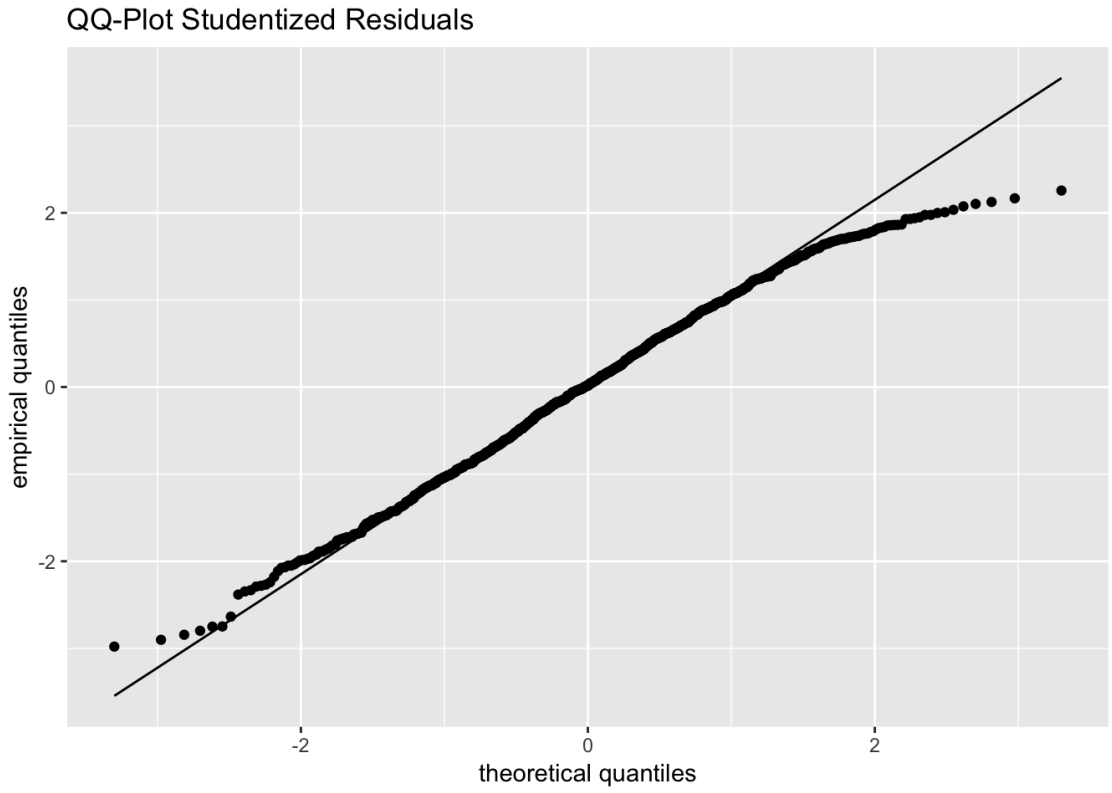
In order to avoid biased estimates and to ensure that any prior findings about the variable effects are accurate, it is necessary to confirm that the model assumptions are respected.

The linear regression model relies on the assumptions that:
1. the error terms are independent random variables
2. the error terms have mean zero
3. the error terms have constant variance (homoscedasticity)
4. the error terms follow a normal distribution

First, to verify the normality of the residuals, a histogram and qq-plot will be used.

```
#Adding the residual and fitted values to the dataset for analysis
resid<-rstudent(mod1a)
fitted<-mod1a$fitted.values
res.dat<-cbind(mydata1,fitted,resid)
head(res.dat)
```

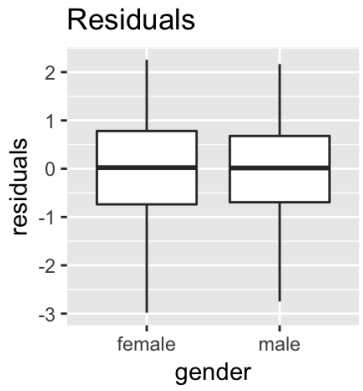


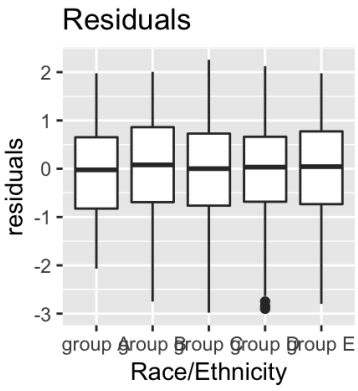


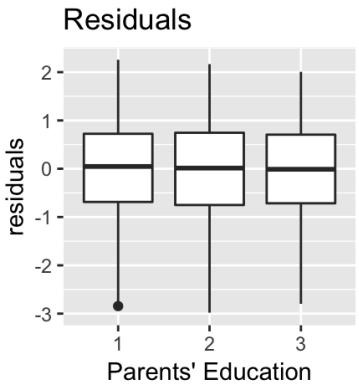
As can be seen from the histogram, the overall shape does resemble a bell-shaped curve. The distribution of the residuals appears to be grouped around the center of the plot, around 0, and is roughly symmetrical on both sides, but slightly skewed to the left.

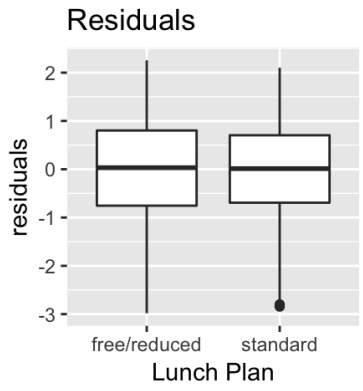
Looking at the qq-plot, if both sets of quantiles came from the same distribution, the qq-plot should show points forming an approximately straight line. It seems that there's good alignment between the theoretical and empirical quantiles except for the tails, where we can observe more deviation. Despite these concerns, the fit does not appear to be too problematic, indicating that the normality assumption of residuals is reasonable.

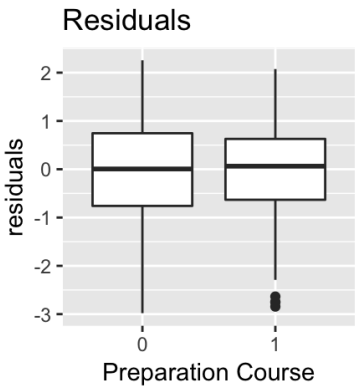
Next, we will create a plot of the residuals against each variable used in the model in order to verify if the model is correctly specify and homogeneity of variance. A boxplot will be generated for `gender`, `race.ethnicity`, `parent_educ`, `lunch` and `prep_course` since they are categorical variables.












```
gender_resid
```

```
## $female
## [1] -5.717746e-05  1.084188e+00
##
## $male
## [1] -0.0001163085  0.9283420942
```

```
race_resid
```

```
## $`group A`
## [1] -2.390737e-05  9.658126e-01
##
## $`group B`
## [1] -4.604429e-05  9.533094e-01
##
## $`group C`
## [1] -2.234919e-05  1.060653e+00
##
## $`group D`
## [1] -0.0001735839  0.9401401086
##
## $`group E`
## [1] -0.0001810893  1.1145202347
```

```
parent_resid
```

```
## $`1`
## [1] -7.095544e-05  9.845511e-01
##
## $`2`
## [1] -6.246884e-05  1.028467e+00
##
## $`3`
## [1] -0.0001830411  0.9922058671
```

```
lunch_resid
```

```
## $`free/reduced`
## [1] -0.000205911  1.129535098
##
## $standard
## [1] -2.467976e-05  9.364891e-01
```

```
prep_resid
```

```
## $`0`
## [1] -0.0001155892  1.0404698480
##
## $`1`
## [1] -3.248183e-05  9.303327e-01
```

The boxplots generated for all variables seem favorable. The spread of the boxplots shows that most values are clustered around the mean, which is close to 0 for all levels of the variables. Moreover, the first and last quartiles are roughly equidistant from the median, suggesting a symmetrical distribution and further supporting the assumption of constant variance.

Given that, the T-test was employed to draw some conclusions earlier, we will also be verifying that the assumption of constant variance among the groups is respected.

```
var.test(avg_score~gender,data=mydata1,alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  avg_score by gender
## F = 1.0693, num df = 482, denom df = 516, p-value = 0.4537
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8972068 1.2751929
## sample estimates:
## ratio of variances
##           1.069346
```

For the variable `gender`, we are testing whether the variance between the two levels (female and male) are equal, that is:
$H_0 : \sigma^2_{genderfemale} = \sigma^2_{gendermale}$
$H_1 : \sigma^2_{genderfemale} \neq \sigma^2_{gendermale}$

We obtain a test statistic of F = 1.0693 (with corresponding df1=482, df2=516) and a p-value of 0.4537, which is greater than any reasonable α. Thus we fail to reject H_0 and conclude that the variances between the two groups are not significantly different. As a result, the assumption of constant variance between the two groups (female and male) seems valid.

```
var.test(avg_score~lunch,data=mydata1,alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  avg_score by lunch
## F = 1.1561, num df = 347, denom df = 651, p-value = 0.1181
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9638883 1.3940131
## sample estimates:
## ratio of variances
##           1.156129
```

For the variable `lunch`, we are testing whether the variance between the two levels (free/reduced and standard lunch plabn) are equal, that is:
$H_0 : \sigma^2_{lunchfree/reduced} = \sigma^2_{lunchstandard}$
$H_1 : \sigma^2_{lunchfree/reduced} \neq \sigma^2_{lunchstandard}$

We obtain a test statistic of F = 1.1561 (with corresponding df1=347, df2=651) and a p-value of 0.1181, which is greater than $\alpha = 0.1$. Thus, at a significance level of $\alpha = 0.1$, we fail to reject H_0 and conclude that the variances between the two groups are not significantly different. As a result, the assumption of constant variance between the two groups (free/reduced and standard lunch plan) seems valid.

```
var.test(avg_score~prep_course,data=mydata1,alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  avg_score by prep_course
## F = 1.0893, num df = 664, denom df = 334, p-value = 0.3754
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9016169 1.3084304
## sample estimates:
## ratio of variances
##             1.0893
```

For the variable `prep_course`, we are testing whether the variance between the two levels (completed vs not completed the preparation course) are equal, that is:
$H_0 : \sigma^2_{PrepCourse0} = \sigma^2_{PrepCourse1}$
$H_1 : \sigma^2_{PrepCourse0} \neq \sigma^2_{PrepCourse1}$

We obtain a test statistic of F = 1.0893 (with corresponding df1=664, df2=334) and a p-value of 0.3754, which is greater than any reasonable α. Thus, we fail to reject H_0 and conclude that the variances between the two groups are not significantly different. As a result, the assumption of constant variance between the two groups (completed vs not completed the preparation course) seems valid.

```
# test: equality of variances (more than 2 levels)
bartlett.test(avg_score~race.ethnicity,data=mydata1)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  avg_score by race.ethnicity
## Bartlett's K-squared = 2.1471, df = 4, p-value = 0.7087
```

For the variable `race.ethnicity`, we are testing whether the variance across the five levels (group A,B,C,D and E) are equal using Barlett test, that is:
$H_0 : \sigma^2_{RaceEthnicityA} = \sigma^2_{RaceEthnicityB} = \sigma^2_{RaceEthnicityC} = \sigma^2_{RaceEthnicityD} = \sigma^2_{RaceEthnicityE}$
H_1 : at least two variances differ

We obtain a test statistic of 2.1471 (with corresponding df=4) and a p-value of 0.7087, which is greater than any reasonable α. Thus, we fail to reject H_0 and conclude that the variances across the groups are not significantly different. As a result, the assumption of constant variance among levels of races/ethnicity seems valid.

```
bartlett.test(avg_score~parent_educ,data=mydata1)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  avg_score by parent_educ
## Bartlett's K-squared = 0.74271, df = 2, p-value = 0.6898
```

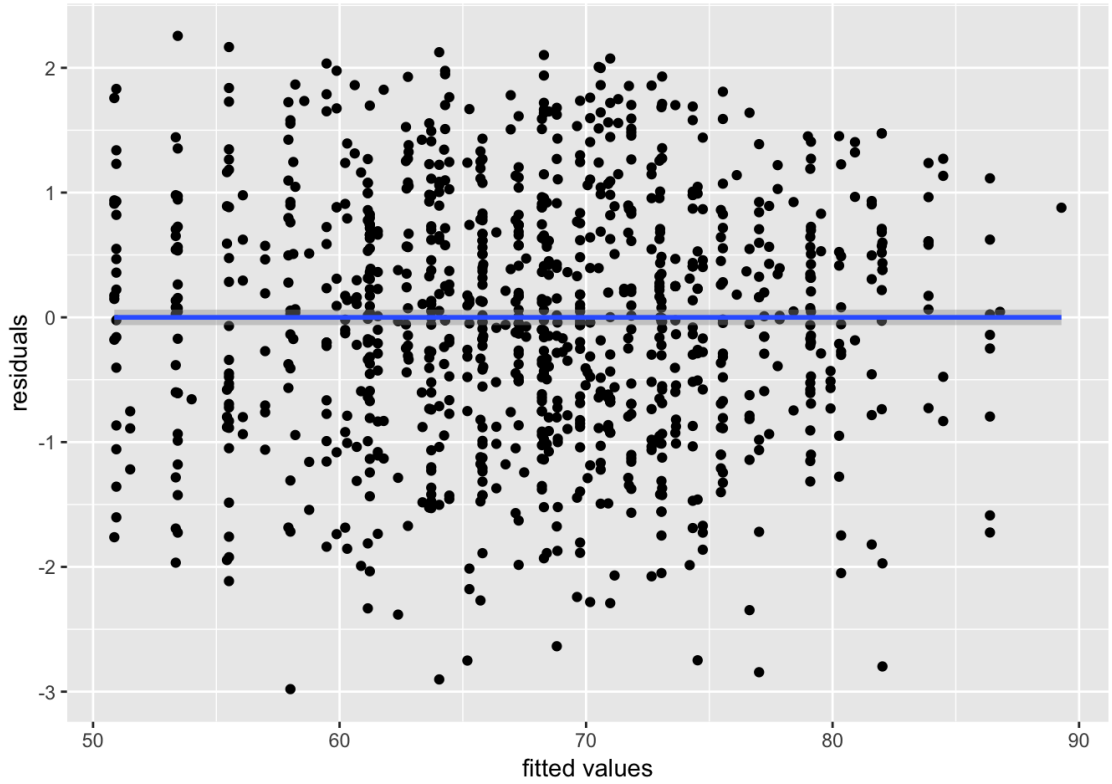
For the variable `parent_educ` , we are testing whether the variance among the three levels (1,2 and 3) are equal using Barlett test, that is:
$H_0 : \sigma^2_{ParentEduc1} = \sigma^2_{ParentEduc2} = \sigma^2_{ParentEduc3}$
H_1 : at least two variances differ

We obtain a test statistic of 0.74271 (with corresponding df=2) and a p-value of 0.6898, which is greater than any reasonable α. Thus, we fail to reject H_0 and conclude that the variances across the groups are not significantly different. As a result, the assumption of constant variance among levels of parents' education seems valid.

Finally, the residuals vs. fitted values of the response variable `avg_score` will be explored.

```
ggplot(data = res.dat,
       aes(x = fitted, y = resid)) +
  geom_point() +
  geom_smooth() +
  theme(legend.position = "bottom") +
  ylab("residuals") +
  xlab("fitted values")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



From the plot, we can observe that the points are roughly symmetrically distributed (around the same magnitude acrosss the x-axis) and appears to be clustered towards the middle of the plot. It is arguable that some trends can be observe around both ends of the plot. However, if we focus on the interval $\hat{y} \in [60, 80]$ where bulk of the data is, the residual plot seems very plausible and does not indicate any obvious signs of heteroscedasticity in general.

In conclusion, the residual analysis findings show that the T-test and assumption of the linear regression model seem to hold, supporting the interpretations and conclusions drawn before.

Question 2

Let's say we decide that a student passes the "exam" if their combined average across all three subjects is higher than 60%. For the sake of simplicity, we will refer to the combination of all three subjects as the "exam." Create an indicator variable `Pass`, which takes value 1 if the student's average score is above 60% and value 0 if below (indicating that he failed). We are now interested to examine the chances that the student will pass the exam.

```
#creating of new variable "Pass"
mydata$pass<-ifelse(mydata$avg_score>=60,1,0)
freqdist(mydata$pass)
```

```
##        frequencies percentage cumulativepercentage
## 0              287       28.7                 28.7
## 1              713       71.3                100.0
## Totals        1000      100.0                100.0
```

With a threshold of 60%, the passing rate will be 71.3% (713 students will pass the exam).

a) Begin by fit a logistic regression that includes only the variable `Prep_Course`***. Use students who have not completed the preparation course as the reference level.***

```
#fitting a logistic regression
mod.log<-glm(pass~prep_course,data=mydata,family=binomial(link="logit"))
summary(mod.log)
```

```
##
## Call:
## glm(formula = pass ~ prep_course, family = binomial(link = "logit"),
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8914  -1.4542   0.6048   0.9238   0.9238
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.63063    0.08144   7.743 9.71e-15 ***
## prep_course1  0.97523    0.16755   5.820 5.87e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1198.9  on 999  degrees of freedom
## Residual deviance: 1161.3  on 998  degrees of freedom
## AIC: 1165.3
##
## Number of Fisher Scoring iterations: 4
```

i) Provide the fitted model on the log-odds scale and probability scale.

- Log-odds scale: The fitted model on the log-odds scale is given by : $ln(\frac{\Pi}{1-\Pi}) = 0.63063 + 0.97523 PrepCourse_1$
  where $\Pi = P(\hat{Pass} = 1|PrepCourse_1)$
- Probability scale:
  The fitted model on the probability scale is given by : where $P(\hat{Pass} = 1|PrepCourse_1) = \frac{exp(0.63063+0.97523 PrepCourse_1)}{1+exp(0.63063+0.97523 PrepCourse_1)}$

ii) Interpret all of the regression coefficients on an appropriate scale. What is the estimated probability of passing the exam when a student has not completed the preparation course? What about for a student who have completed the preparation course?

```
mod.log$coefficients
```

```
##  (Intercept) prep_course1
##    0.6306268    0.9752333
```

```
exp(mod.log$coefficients)
```

```
##  (Intercept) prep_course1
##     1.878788     2.651786
```

- $\hat{\beta}_0 = 0.6306$ or $exp(\hat{\beta}_0) = 1.8788$: When the variable $PrepCourse_1 = 0$ (i.e. a student has not completed the preparation course), the intercept $\hat{\beta}_0 = 0.6306$ represents the estimated log of the odds of having the outcome Pass=1 (by obtaining a score above 60% in the exam or equivalently passing the exam). It is more meaningful to interpret this result on the odds ratio scale by exponentiating $exp(\hat{\beta}_0) = 1.8788$ and say that, the odds of passing the exam when the student has not completed the preparation course is 1.8788.
- $\hat{\beta}_1 = 0.97523$ or $exp(\hat{\beta}_1) = 2.6518$: The interpretation of regression coefficient $\hat{\beta}_1$ is done with respect to the reference level (students who have not completed the preparation course) on the log odds ratio scale. It is more meaningful to interpret this result on the odds ratio scale by exponentiating the estimated coefficient $exp(\hat{\beta}_1) = 2.6518$ and say that, the odds of passing the exam for a student who have completed the preparation course is 2.6518 times the odds of passing the exam for a student who have not completed the preparation course. Given that this ratio>1, it suggests that the odds are higher for students who have completed the preparation course vs. those who have not.

The estimated probabilities of passing the exam for students who have not completed the preparation course $PrepCourse_1 = 0$ is 62.53%.

$$ln(\frac{P(\hat{Pass}=1|PrepCourse_1=0)}{1-P(\hat{Pass}=1|PrepCourse_1=0)}) = 0.6306$$
$$\frac{P(\hat{Pass}=1|PrepCourse_1=0)}{1-P(\hat{Pass}=1|PrepCourse_1=0)} = e^{0.6306} = 1.8788$$
$$P(\hat{Pass} = 1|PrepCourse_1 = 0) = \frac{1.8788}{1+1.8788} = 0.62526$$

The estimated probabilities of passing the exam for students who have completed the preparation course $PrepCourse_1 = 1$ is 83.28%.

$$ln(\frac{P(\hat{Pass}=1|PrepCourse_1=1)}{1-P(\hat{Pass}=1|PrepCourse_1=1)}) = 0.6306 + 0.9752 = 1.606$$
$$\frac{P(\hat{Pass}=1|PrepCourse_1=1)}{1-P(\hat{Pass}=1|PrepCourse_1=1)} = e^{1.606} = 4.9821$$
$$P(\hat{Pass} = 1|PrepCourse_1 = 1) = \frac{4.9821}{1+4.9821} = 0.8328$$

iii) A student is unsure if enrolling in the preparation course will actually improve his chances of passing the test. Based on the results, what would you recommend? Support your explanation with a statistical test.

We are interested to test whether β_1, that is, the coefficient representing the log of the odds ratios between $PrepCourse_1 = 1$ vs $PrepCourse_1 = 0$ (the reference level) is significant in the model. Formally, the hypothesis we are testing are :
$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

From the output summary, the result of the Wald's test statistic is z=5.820 with a corresponding p-value of 5.87e-09, which is < $\alpha = 0.001$. Therefore, we reject H_0 in favor of H_1 and conclude that, at any reasonable level of α, the log of the odds ratios between a student who has completed the preparation course and a student who has not completed the preparation course is significantly different from 0. This implies that the odds ratios must also be significantly different from 0, so we say state that having completed the preparation course does have a significant effect on the odds of passing the exam.

Combining this result with the large effect size found earlier in ii), we can observe that enrolling in the preparation course does increase the probability of passing the exam. More specifically, it increases the odds by 2.6518. Therefore, we would recommend the student to enroll in this preparation course.

b) Suppose that now we are interested to investigate the number of subject that a student passes, using the same threshold of 60%. Create an indicator variable `total_pass` ***, which counts the number of subjects in which he receives a grade of 60% or more. We are now interested to examine the number of school subjects each student passes***

```
count<-raw_data
count$pass_math<-ifelse(raw_data$math.score>=60,1,0)
count$pass_reading<-ifelse(raw_data$reading.score>=60,1,0)
count$pass_writing<-ifelse(raw_data$writing.score>=60,1,0)
count$count_pass<-count$pass_math+count$pass_reading+count$pass_writing
mydata$total_pass<-count$count_pass
freqdist(mydata$total_pass)
```

```
##        frequencies percentage cumulativepercentage
## 0              192       19.2                 19.2
## 1               75        7.5                 26.7
## 2              145       14.5                 41.2
## 3              588       58.8                100.0
## Totals        1000      100.0                100.0
```

Looking at the distribution of the new variable `total_pass` , we can observe that 19.2% of students failed all three subjects. There are 7.5% and 14.5% of students who passed only one subject and two subjects respectively. Finally, the proportion of students who passed all three subjects is 58.8%.

i) Fit a Poisson regression model using the variables `Gender` , `Race.Ethnicity` , `Parent_Educ` , `Lunch` ***and*** `Prep_course` . ***Provide the fitted model on the mean scale.***

```
mod.poi<-glm(total_pass~gender+race.ethnicity+parent_educ+lunch+prep_course,data=mydata,family=poisson(link
="log"))
summary(mod.poi)
```

```
##
## Call:
## glm(formula = total_pass ~ gender + race.ethnicity + parent_educ +
##     lunch + prep_course, family = poisson(link = "log"), data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4397  -0.5953   0.1887   0.5282   1.3504
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.28927    0.09681   2.988 0.002809 **
## gendermale            -0.05418    0.04381  -1.237 0.216269
## race.ethnicitygroup B -0.03204    0.09295  -0.345 0.730307
## race.ethnicitygroup C -0.02786    0.08808  -0.316 0.751729
## race.ethnicitygroup D  0.17203    0.08799   1.955 0.050589 .
## race.ethnicitygroup E  0.21965    0.09703   2.264 0.023589 *
## parent_educ2           0.13295    0.04924   2.700 0.006930 **
## parent_educ3           0.21146    0.06017   3.514 0.000441 ***
## lunchstandard          0.37025    0.04928   7.513 5.79e-14 ***
## prep_course1           0.21032    0.04468   4.707 2.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1060.36  on 999  degrees of freedom
## Residual deviance:  938.16  on 990  degrees of freedom
## AIC: 3246.4
##
## Number of Fisher Scoring iterations: 5
```

The fitted model is given by :

$$\hat{E(Total\_pass}|Gender_{male}, Race.Ethnicity_B, Race.Ethnicity_C, Race.Ethnicity_D,$$
$$Race.Ethnicity_E, ParentEduc_2, ParentEduc3, Lunch_{standard}, PrepCourse_1)$$
$$= \hat{\lambda} = exp(0.28927 - 0.05418Gender_{male} - 0.03204Race.Ethnicity_B - 0.02786Race.Ethnicity_C$$
$$+0.17203Race.Ethnicity_D + 0.21965Race.Ethnicity_E + 0.13295ParentEduc_2 + 0.21146ParentEduc3$$
$$+0.37025Lunch_{standard} + 0.21032PrepCourse_1)$$

ii) Give an interpretation of the intercept in model.

The intercept $exp(\beta_0)$ represents the mean of the response variable when all explanatory variables are zeros. In this context, the intercept $exp(0.28927) = 1.3355$ is the expected number of subjects passed for a student who is female, from race/ethnicity A, with parents who have completed high school (or some high school), enrolled in the free/reduced lunch plan and has not completed the preparation course.

iii) Assess the global significance of the variable `Parent_Educ` ***in the model.***

```
library(car)
Anova(mod.poi, type=3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: total_pass
##                LR Chisq Df Pr(>Chisq)
## gender            1.529  1   0.216318
## race.ethnicity   24.005  4  7.969e-05 ***
## parent_educ      14.028  2   0.000899 ***
## lunch            59.432  1  1.266e-14 ***
## prep_course      21.773  1  3.068e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We are interested to test the global significance of the variable Parent_Educ, that is, testing whether the coefficients associated with all the levels of Parent_Educ is equal to 0, or at least some coefficients are significantly other than 0.

$H_0 : \beta_{ParentEduc2} = \beta_{ParentEduc3} = 0$
H_1 : at least one of $\beta_{ParentEduc2}$ or $\beta_{ParentEduc3} \neq 0$

The results from Anova is based on a χ^2 distribution with 2 degree of freedom. We can observe that the test ChiSq=14.028 with a corresponding p-value of 0.000899, which is $< \alpha = 0.001$. Therefore, we reject H_0 in favor of H_1 and conclude that, at least one of $\beta_{ParentEduc2}$ or $\beta_{ParentEduc3} \neq 0$. Thus, the variable Parent_Educ is globally significant in the model containing gender, race.ethnicity, lunch and prep_course. In other words, the parents' education level does have an effect on the mean number of subjects passed by the student, holding all other variables constant.

iv) Discuss in a few sentences the main assumptions differences between a Poisson and Quasi-Poisson model. What are the benefits or drawbacks of using a Quasi-Poisson model instead of the Poisson model.

The Poisson model makes the assumption that the variance is equal to the mean, which makes this model quite restrictive and not always a fair assumption in real-life application, since not all distributions may adhere to this relationship between the variance and the mean. The situation where the variance is greater than the mean is known as Overdispersion (in a Poisson model, the overdispersion parameter=1) and it can be handled with a Quasi-Poisson model. Since a Quasi-Poisson model assumes that the variance is a linear function of the mean, it is a more suitable model to employ for a count response variable when the variance is greater than the mean.

The benefits of a Quasi-Poisson model are its flexibility and its ability to handle overdispersion. It will result in the same estimates for the regression parameters and will also provide a better fit to the data. However, the drawbacks is that this approach is based on a pseudo-likelihood, meaning that the underlying equations used to estimate is not based on a likelihood function from a proper distribution, but rather a "quasi" likelihood function, thus some properties, indices or tests specific to MLE cannot be applied.

c. Repeat b) with a Negative Binomial regression model this time.

```
library(MASS)
mod.nb<-glm.nb(total_pass~gender+race.ethnicity+parent_educ+lunch+prep_course,data=mydata)
summary(mod.nb)
```

```
##
## Call:
## glm.nb(formula = total_pass ~ gender + race.ethnicity + parent_educ +
##     lunch + prep_course, data = mydata, init.theta = 55433.90366,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4397  -0.5953   0.1887   0.5282   1.3504
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.28926    0.09681   2.988 0.002809 **
## gendermale            -0.05418    0.04382  -1.236 0.216275
## race.ethnicitygroup B -0.03204    0.09295  -0.345 0.730317
## race.ethnicitygroup C -0.02786    0.08808  -0.316 0.751740
## race.ethnicitygroup D  0.17203    0.08800   1.955 0.050592 .
## race.ethnicitygroup E  0.21965    0.09703   2.264 0.023590 *
## parent_educ2           0.13295    0.04924   2.700 0.006930 **
## parent_educ3           0.21146    0.06017   3.514 0.000441 ***
## lunchstandard          0.37025    0.04928   7.513 5.79e-14 ***
## prep_course1           0.21032    0.04468   4.707 2.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(55433.9) family taken to be 1)
##
##     Null deviance: 1060.34  on 999  degrees of freedom
## Residual deviance:  938.14  on 990  degrees of freedom
## AIC: 3248.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  55434
##           Std. Err.:  229896
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -3226.369
```

i) Compare the Poisson model to the Negative Binomial using a proper statistical test. Would the Poisson model be an adequate simplification of the Negative Binomial model?

```
-2*as.numeric(logLik(mod.poi))
```

```
## [1] 3226.352
```

```
-2*as.numeric(logLik(mod.nb))
```

```
## [1] 3226.369
```

```
lrtstat <- -2*as.numeric(logLik(mod.poi)-logLik(mod.nb))
lrtstat
```

```
## [1] -0.01722744
```

```
pchisq(lrtstat, df = 1, lower.tail = FALSE)/2
```

```
## [1] 0.5
```

To test whether the Poisson model is an adequate simplification of the Negative Binomial model, we have to test the following: $H_0 : k = 0$ vs $H_1 : k > 0$ since the negative binomial model simplifies to the Poisson model when the parameter k=0.

The result of the LRT = -0.01722744 with an associated p-value = 0.5. Given that this p-value is greater than any reasonable α, we fail to reject the null hypothesis that $H_0 : k = 0$. Thus, the result of this test statistic suggests that the Poisson model is indeed an adequate simplification of the Negative Binomial model.

ii) Use the AIC and BIC criterion to compare the Poisson model and Negative Binomial now. Which model is selected by each criterion?

```
AIC(mod.poi)
```

```
## [1] 3246.352
```

```
AIC(mod.nb)
```

```
## [1] 3248.369
```

```
BIC(mod.poi)
```

```
## [1] 3295.43
```

```
BIC(mod.nb)
```

```
## [1] 3302.355
```

Under both approach, a smaller value indicates a better model.
Under the AIC, Poisson: 3246.352 < Negative Binomial: 3248.369
Under the BIC, Poisson: 3295.43 < Negative Binomial: 3302.355

From the results, we can observe that the Poisson model is selected according to both criteria and is therefore selected as the better model.

Question 3

Consider the distribution Weibull $f(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} exp(-(\frac{x}{\lambda})^k), where\ x \geq 0; \lambda, k > 0$

a. Write an expression for the likelihood function.

for sample $x_1, x_2, \ldots, x_n$ the likelihood function is:

$$\begin{aligned} L(\lambda, k) &= \prod_{i=1}^{n} f(x_i; \lambda, k) \\ &= \prod_{i=1}^{n} \frac{k}{\lambda}(\frac{x_i}{\lambda})^{k-1} exp(-(\frac{x_i}{\lambda})^k) \\ &= (\frac{k}{\lambda^k})^n exp(-\sum_{i=1}^{n} \frac{x_i^k}{\lambda^k}) \prod_{i=1}^{n} x_i^{k-1} \end{aligned}$$

b. Write an expression for the log-likelihood function.

$$
\begin{aligned}
LL(\lambda, k) &= \ln[\prod_{i=1}^{n} f(x_i)] \\
&= \ln[(\frac{k}{\lambda^k})^n exp(-\sum_{i=1}^{n} \frac{x_i^k}{\lambda^k}) \prod_{i=1}^{n} x_i^{k-1}] \\
&= n \ln k - nk \ln \lambda - \sum_{i=1}^{n} \frac{x_i^k}{\lambda^k} + (k-1) \sum_{i=1}^{n} \ln x_i
\end{aligned}
$$

c. What is the maximum likelihood estimator for k, λ?

For λ:

$\frac{\partial \ln L(k,\lambda)}{\partial \lambda} = -\frac{nk}{\lambda} + \frac{k \sum_{i=1}^{n} x_i^k}{\lambda^{k+1}}$

let $\frac{\partial \ln L(k,\lambda)}{\partial \lambda} = 0$, it follows:

$$
\begin{aligned}
-\frac{nk}{\lambda} + \frac{k \sum_{i=1}^{n} x_i^k}{\lambda^{k+1}} &= 0 \\
\frac{\sum_{i=1}^{n} x_i^k}{n} &= \lambda^k \\
\Rightarrow \lambda^* &= (\frac{\sum_{i=1}^{n} x_i^{k^*}}{n})^{\frac{1}{k^*}}
\end{aligned}
$$

For k:

$\frac{\partial \ln L(k,\lambda)}{\partial k} = \frac{n}{k} - n \ln \lambda - \sum_{i=1}^{n} (\frac{x_i}{\lambda})^k \ln(\frac{x_i}{\lambda}) + \sum_{i=1}^{n} \ln x_i$

let $\frac{\partial \ln L(k,\lambda)}{\partial k} = 0$ and plug in $\lambda^* = (\frac{\sum_{i=1}^{n} x_i^{k^*}}{n})^{\frac{1}{k^*}}$

we could have:

$k^* = [\frac{\sum_{i=1}^{n} x_i^{k^*} \ln x_i}{\sum_{i=1}^{n} x_i^{k^*}} - \frac{\sum_{i=1}^{n} \ln x_i}{n}]^{-1}$