STA 106 Group Project2

By:Han Zhou,Yingyue Chen, Tingwei Zhang

2024-11-23

STA 106 Group Project2

Topic 1

By:Han Zhou,Yingyue Chen, Tingwei Zhang

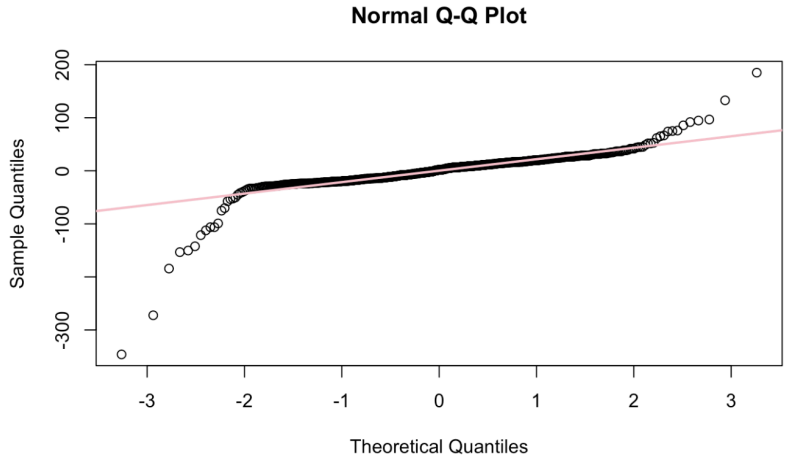
2024-11-23

I-Introduction

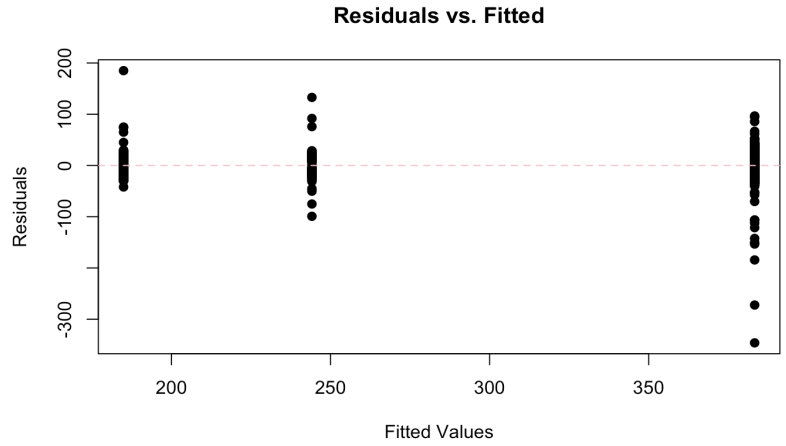
This study analyzes the wing length data of 907 hawks, aiming to determine whether wing length can effectively distinguish different species of hawks. The study includes three types of hawks: Cooper's Hawks, Red-tailed Hawks, and Sharp-shinned Hawks. For this study, we will remove the outliers and transform the given data to fit the ANOVA model. Moreover, we will explain our results and provide suggestions. Our goal is to compare the wing lengths of these species to identify distinguishing features and provide a scientific basis for species classification.

II-Original data diagnostic & model fit

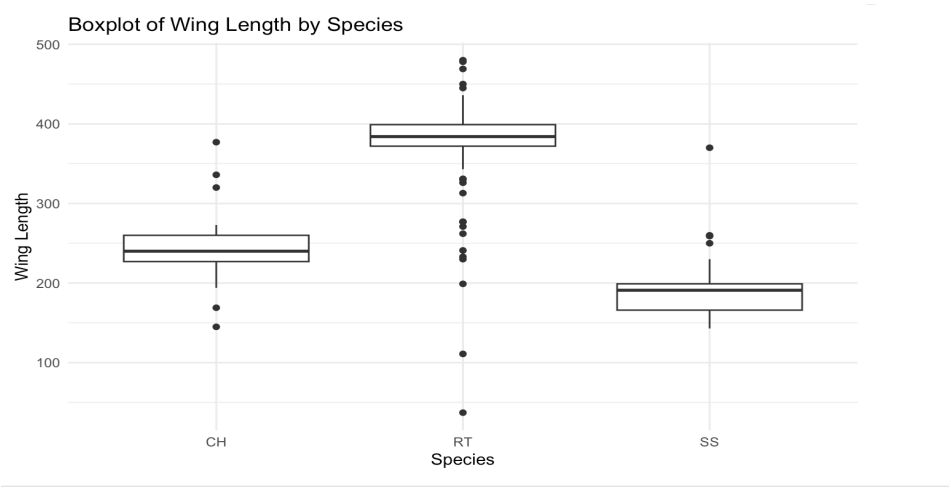
1) Diagnostic plots



Most of the data points closely follow the diagonal line, indicating that the data roughly follows a normal distribution. However, the tails deviate significantly, suggesting the presence of outliers.



Residuals are centered around zero. However, the spread of residuals seems uneven, particularly with more variability on the right side, which suggests potential heteroscedasticity.



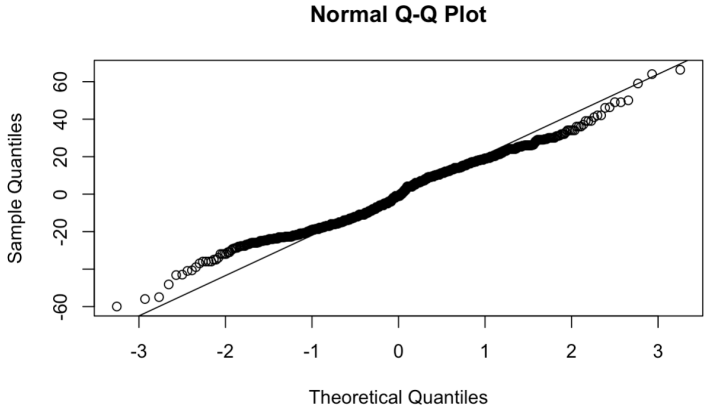
This boxplot shows the wing lengths for three hawk species: Cooper's Hawks (CH), Red-tailed Hawks (RT), and Sharp-shinned Hawks (SS). Red-tailed Hawks have the longest wings, Sharp-shinned Hawks have the shortest, and Cooper's Hawks are in between. Red-tailed Hawks (RT) have outliers with longer wings, suggesting higher variability. Wing length appears useful for distinguishing species, but statistical analysis is needed to confirm the significance.

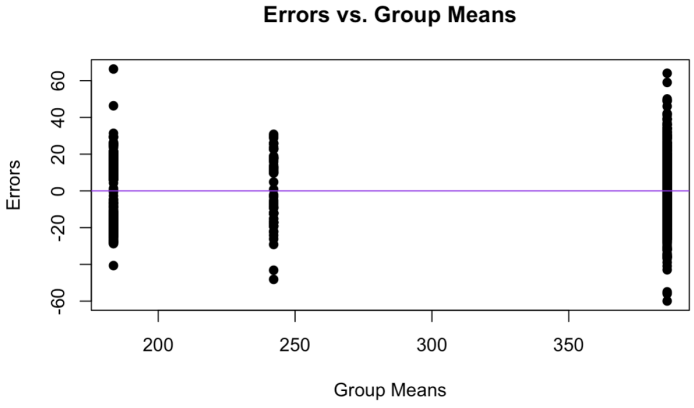
2) Original Model Fit

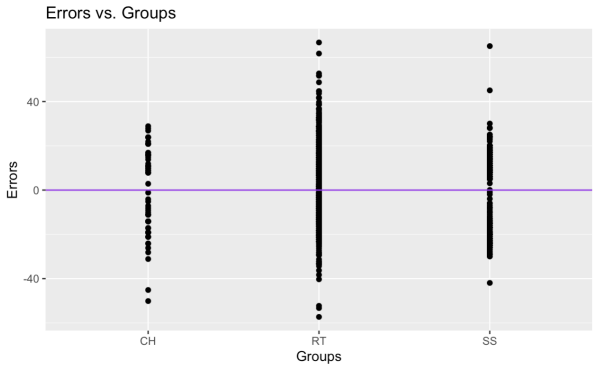
The original ANOVA model indicates that Species has a significant impact on the Wing variable, with an R-square of 0.9061, meaning that Species can explain approximately 90.61% of the Wing. The minimum value of the residual is -346.10 and the maximum value is 185.05, indicating a large range of variation. Coefficient analysis shows that SpeciesRT has a positive effect on Wing (139.159), while SpeciesSS has a negative effect on Wing (-59.199). Additionally, the F-statistic of 4363 (with degrees of freedom 2 and 904) and a p-value of less than 2.2e-16 confirm the model is highly statistically significant overall. Although the overall fitting effect of the model is good, residual plots analysis shows that there is a bias in the tail. So we need to remove outliers or transform the data to improve the model.

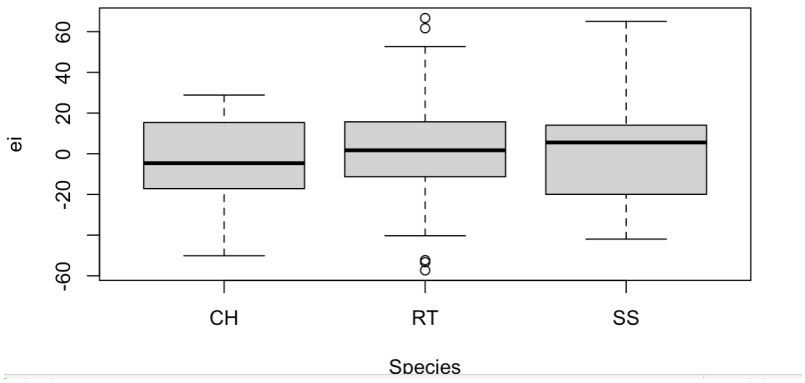
III-outlier removal and transformation

1) outlier removal & plots







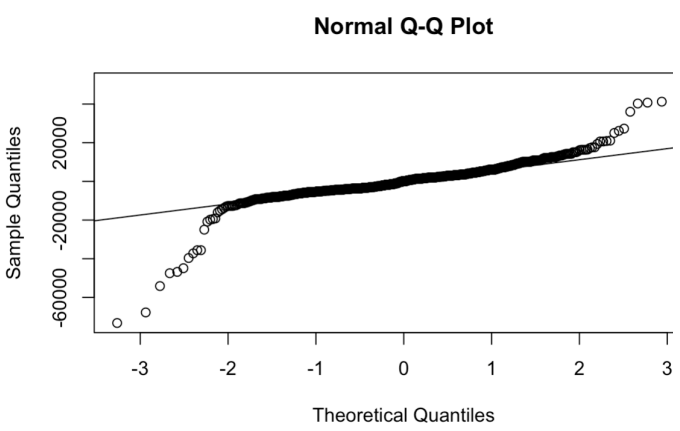


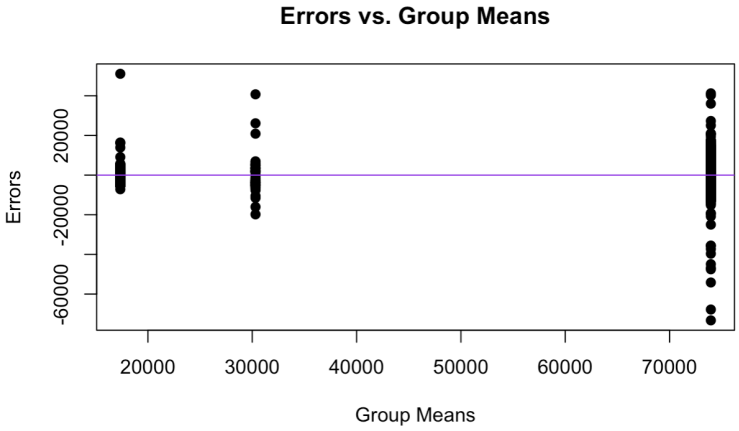
This is the result after removing outliers, and it can be seen that the overall performance of the model has improved. Firstly, the QQ chart shows that most of the residual points fall on the diagonal, indicating that the residuals roughly follow a normal distribution. But there is still a slight deviation at the tail, especially in the negative direction. The scatter plot of residuals and fitted values shows that residuals are generally uniformly distributed around zero, and there is no obvious systematic pattern. In each group, the center of the residuals is concentrated around zero, but in some groups (such as RT), the diffusion of residuals is larger. The box plot also shows that the residuals of CH and SS are relatively concentrated, while the residual range of the RT group is slightly larger and has some outliers.

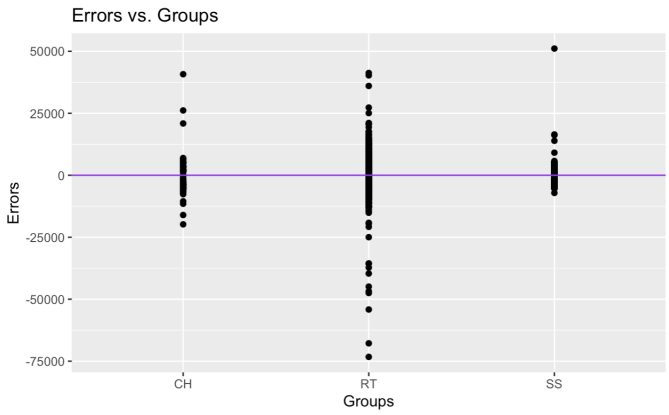
Shapiro-Wilk (normality test)	6.14e-08
Brown-Forsythe (equal variance test)	0.2262858

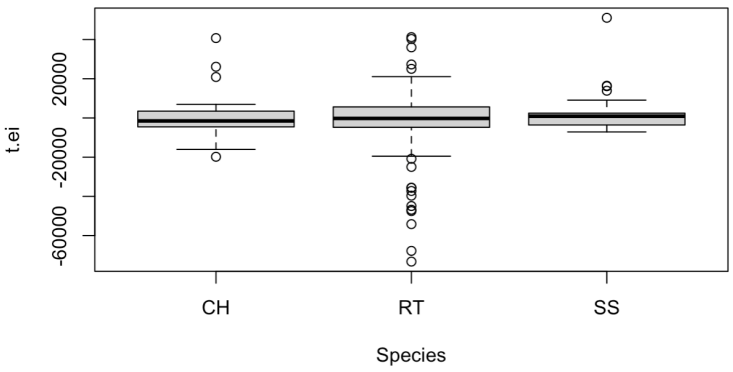
We conducted another formal test to verify normality and equal of variance. The Shapiro Wilk normality test results showed a p-value of 6.14e-08, far below the significance level of 0.05, indicating a significant deviation from the normal distribution of the data. On the other hand, the p-value of Levene's equal of variance test is 0.2263, which is greater than 0.05, indicating that the variances of each group of data can be considered equal and meet the hypothesis of ANOVA analysis.

2) transformation & plots









We used Box Cox transformation and optimized the original model based on different methods (PPCC, Shapiro Wilk, and Log Likelihood). By comparing the results of each method, lambda=2 was ultimately chosen as the transformation parameter to make the data more in line with the normality assumption and improve the fitting effect of the model.

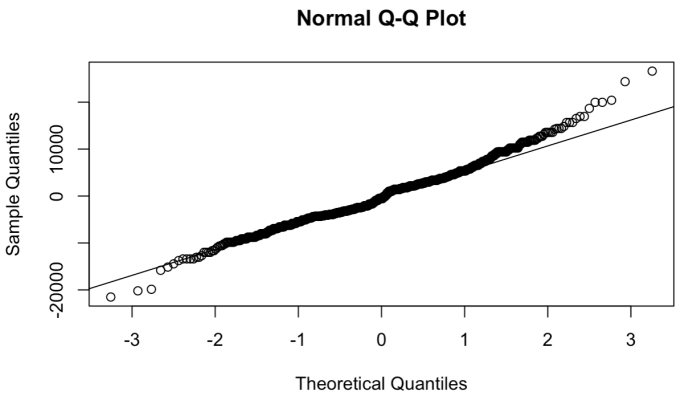
From the QQ plot, it can be seen that most of the residual points fall near the diagonal, but the difference from the original data is very small, and there is still a significant deviation in the tail region. The scatter plot of error and group means shows that the residuals are uniformly distributed around zero values, but

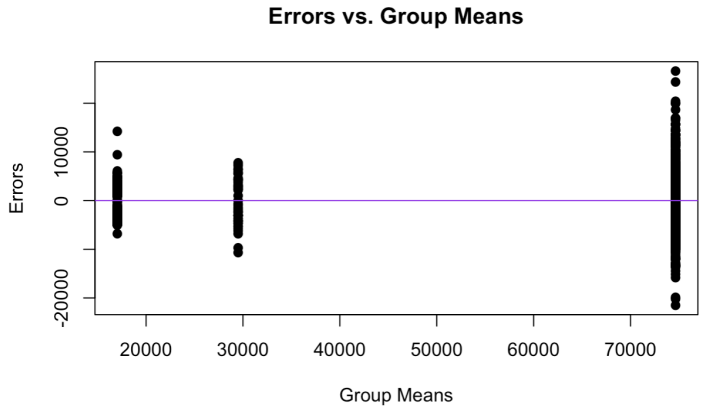
there are significant fluctuations and outliers in the high fitted value regions. In the error vs. groups, the residuals of CH and SS groups are more concentrated, while the residuals of the RT group fluctuate greatly and have obvious outliers, which is also reflected in the box plot.

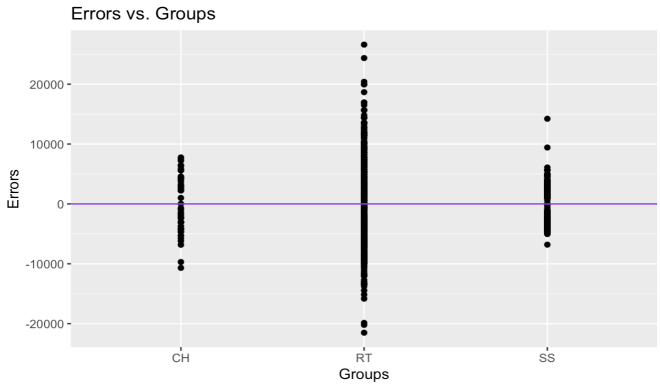
Shapiro-Wilk (normality test)	2.2e-16
Brown-Forsythe (equal variance test)	2.783723e-11

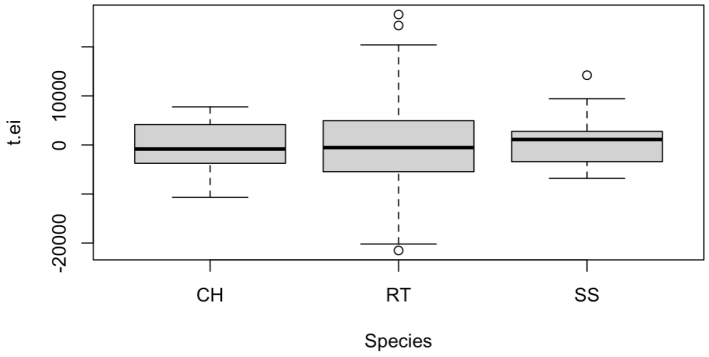
The results of the Shapiro Wilk normality test show that the p-value is less than 2.2e-16, which further indicates that the data deviates significantly from the normal distribution. In addition, the p-value of Levene's homogeneity of variance test is 2.78e-11, significantly less than 0.05, indicating that the variances between the groups are unequal and do not meet the assumption of homogeneity of variance.

3) outlier removal and transformation & plots









After transforming the outlier removed data, we get the new QQ plot, it can be seen that the residual distribution after transformation is relatively closer to normal, indicating an improvement in the normality of the data. The scatter plot of error and group means shows that most residuals are uniformly distributed around zero, but there are some outliers in areas with larger fitted values. In group analysis, the residuals of CH and SS groups are relatively concentrated, but the residuals of the RT group are more scattered and have more outliers, which is confirmed in the group residual graph and box plot. The RT group showed a larger residual range and more outliers.

Shapiro-Wilk (normality test)	6.837e-08
Brown-Forsythe (equal variance test)	2.00765e-23

In terms of statistical testing, the p-value of the Shapiro Wilk normality test is 6.837e-08, which is much lower than the significance level of 0.05, indicating that the residuals deviate significantly from the normal distribution. The p-value of Levene's homogeneity of variance test is 2.00765e-23, which is also less than 0.05, indicating that the inter group variance is not homogeneous and the homogeneity of variance hypothesis is not satisfied. Therefore, although the transformation has improved normality to some extent, there are still biases in the residual distribution, and the heterogeneity of variance may affect the applicability of the model.

IV-Discuss result

On one hand, the different ways to transform have different help on the data. First, it can improve the normality, Second, it can reduce the extreme variability to make the group variance equal. Finally, it can remove the outlier to reject the bias of values in given data.In this case study, removing outliers alone reduces the variability to meet the assumption of ANOVA analysis. However, the normality is decreasing. For both transformation dependent variables and combining transformation dependent variables and removing outliers, they will not only decrease the normality, but also have unequal variance between groups. On the other hand, the downside of transformation is significant. Transformation will be more difficult to interpret due to the change in the unit of the variable. Also, it is difficult to convert them back to the original data after transformation.

In this case study, transformation data is not better fit and didn’t meet the assumption of ANOVA. Only removing the outliers can improve the group variance equal, but it decreases the normality. For both

transformation dependent variable and combine transformation dependent variable and remove outlier cause a violation in the assumption of equal variance and the assumption of normality of errors. Nevertheless, the transformation data is better fit in general, removing outlier and transformation allows the data to become more normal and make the group variance more equal to meet ANOVA assumption. We can first determine the result by using QQ plot and scatter plot of error and group means. However, those plots are subjective and difficult to interpret. As a result, we need a more specific and straightforward way to test whether the data meet the ANOVA assumption, such as Shapiro-Wilk (normality test),Brown-Forsythe (equal variance test).

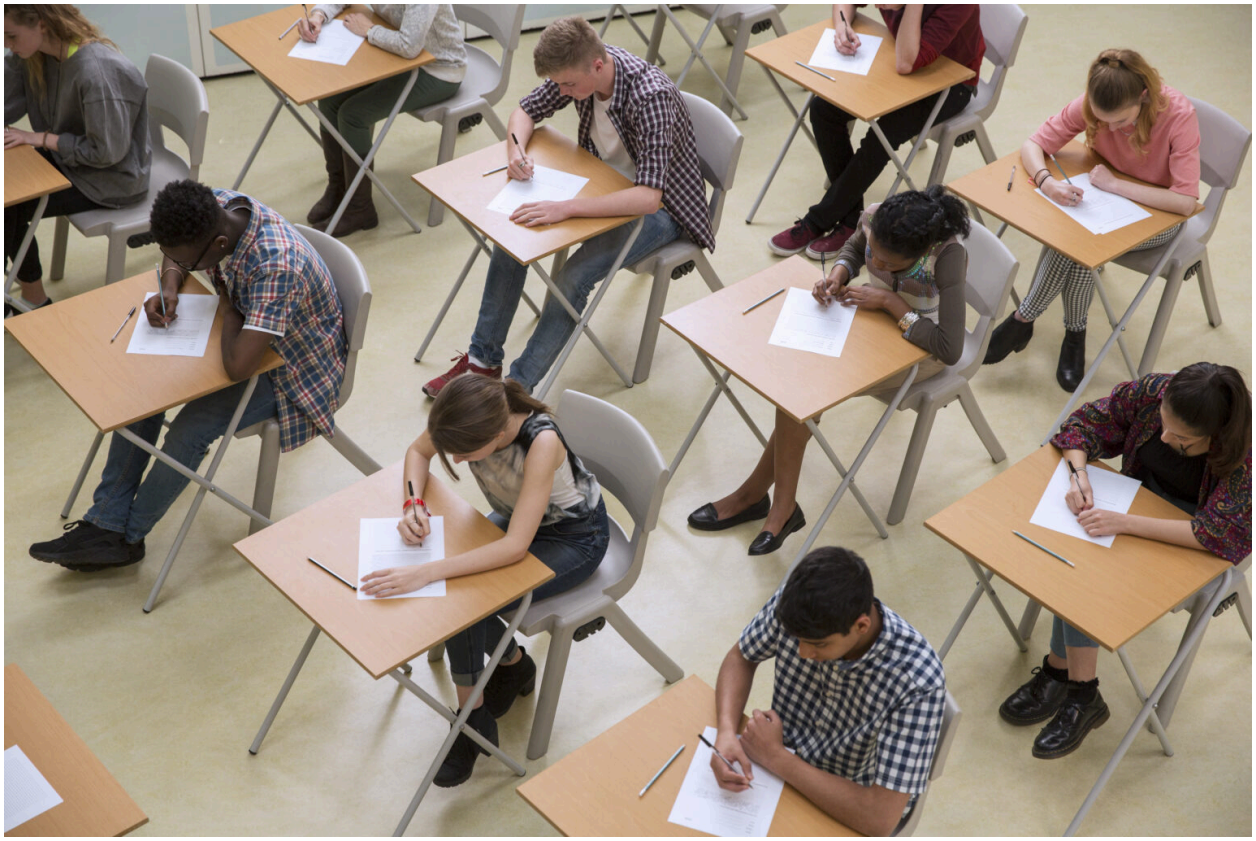
Finally, we will give some suggestions for a client who wants to use this data set for ANOVA. We recommend the client remove the outliers and not use the transformation. Removing outliers can improve normality. Outliers have a significant impact on the calculated values of the data, we can remove the outlier when the proportion of outliers in the sample size is small. However, we don’t recommend doing any transformation or combined transformation and outlier removal. This is because it can’t improve normality and make the group variance equal to meet the ANOVA assumption.

STA 106 Group Project2

Topic 2

By:Han Zhou,Yingyue Chen,Tingwei Zhang

2024-11-23



I-Introduction

The purpose of this study is to determine whether students' performance is affected by gender and/or their parents' education level. We focus on gender and Parental Level of Education as the main explanatory factors and use average student scores as the response variable. Gender refers to whether the student is male or female, and parental education refers to associate's degree, bachelor's degree, high school, master's degree, some college, some high school. Identifying these differences is vital for schools to increase the overall class performance and ensure that every student has equal opportunity to access education. To solve this problem, we apply two-way Factor ANOVA to determine whether the interaction effect of gender and parental level of education is significant by comparing Reduced Model and Full Model. Finally, we analyze pairwise confidence intervals for the main effect to gain a deeper understanding on specific differences within each factor and the significance of differences between different groups.

II-Summary of the data

1) Summary statistic

Parental Level of Education gender	associate's degree	bachelor's degree	high school
male(1)	67.86792 (SD=13.86722,n=106)	68.77576 (SD=13.06380,n=55)	61.57516 (SD=12.78029,n=102)
Female(0)	71.12356 (SD=13.35958,n=116)	74.67196 (SD=14.21084,n=63)	64.74823 (SD=14.14405,n=94)

Parental Level of Education gender	master's degree	some college	some high school
male(1)	73.52174 (SD=11.99656,n=23)	65.71605 (SD=14.21565,n=108)	64.63636 (SD=13.26935,n=88)
Female(0)	73.64815 (SD=14.69973,n=36)	71.00282 (SD=12.77644,n=118)	65.56410 (SD=16.53546,n=91)

This table shows a summary of the data. It illustrates the means (average scores of average scores that each student's performance in math, reading, and writing), standard deviations (the spread of the data collected for each level of education and each gender), and the sample size (the number of individuals in each level of education and each gender).

The first column represents gender—male and female—and the first row represents the different parental levels of education: associate's degree, bachelor's degree, high school, master's degree, some college, and some high school. The statistics are presented inside the table. For example, for male students whose parental level of education is an associate's degree, the average performance score is 67.86792 (the average score of average score of math, reading, and writing performance for these students), the standard deviation is 13.86722 (the variability in performance scores for this group), and the sample size is 106 (the number of male students whose parental level of education is an associate's degree).

The treatment sample sizes indicate that we are dealing with an unbalanced design, where some treatment sizes are significantly smaller than others. For example, there are only 23 male students whose parental level of education is a master's degree, but there are 118 female students whose parental level of education is some college.

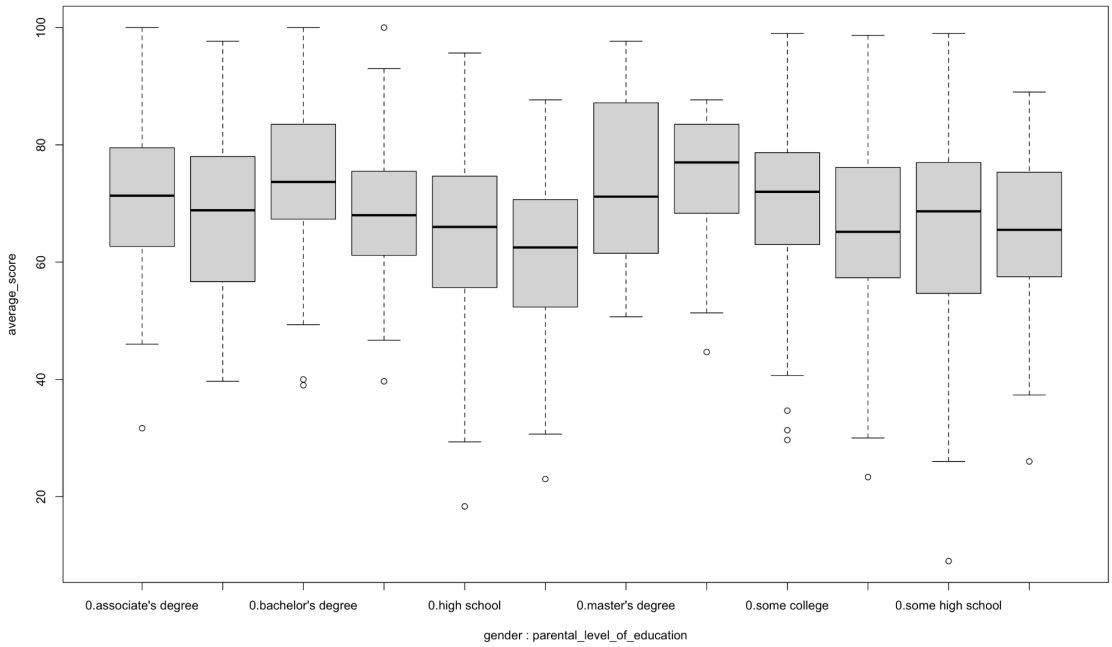
Based on the mean, we can determine the average scores of average scores that each student's performance in math, reading, and writing, gender differences and parental

education level both affect the average scores. Female students have higher average scores than the male students, and a student's parent with higher level of education will have higher scores than the student's parent with lower level of education. For males, the highest average score is 73.52174 (male students,parental level of education is master's degree), lowest average scores is 61.57516(male students, parental level of education is high school). For females, the highest average score is 74.67196(female students,parental level of education is bachelor's degree), the lowest average score is 64.74823(female students, parental level of education is high school)

Based on the standard deviation, we can speculate that the data set has a similar(nearly equal) variance which meets the assumption of the ANOVA. For example, in the students whose parental level of education is associate's degree, male student standard deviation is 13.86722 and female student standard deviation is 13.35958. The lowest standard deviation is 11.99656(male student,parental level of education is master's degre), the highest standard deviation is 16.53546(female students,parental level of education is high school). There are differences in standard deviation, but it is not too significant. However, it is only speculation, we need to use more specific statistical test to determine whether the data set has an equal variance that meets the assumption of ANOVA.

We speculate there may not be interaction between gender and parental level of education, because all genders follow the similar trend dependent on the parental level of education. However, there may be the main effects for both gender and parental level of education.For example, in the change from "master's degree" to "some college", for males, it have a significant decreasing(73.52174 to 65.71605), but for females, it only have little decreasing(73.64815 to 71.00282). The trend of them is all decreasing, but there might have different levels of decrease which lead to a little interaction between gender and parental level of education. However, it is only the speculation, we need use more specific statistical tests to determine the interaction between gender and parental level of education.

Boxplot



The boxplot illustrates the median,variability, quartiles and outliers of each region group. For each degree, the left one represents female, and the left one represents male.

Overall, female students have higher average scores than the male students, and a student's parent with a higher level of education will have higher scores than the student's parent with a lower level of education. Female students with bachelor’s degree parent have highest median (highest average scores) followed by female students with master’s degree parents, male students with master’s degree parents, female students with associate’s degree parents, female students with some college parents, male students with bachelor’s degree parents, male students with associate’s degree parents, male students with some college parents, female students with some high school parents, female students with high school parents, male students with some high school parents and male students with high school parents.

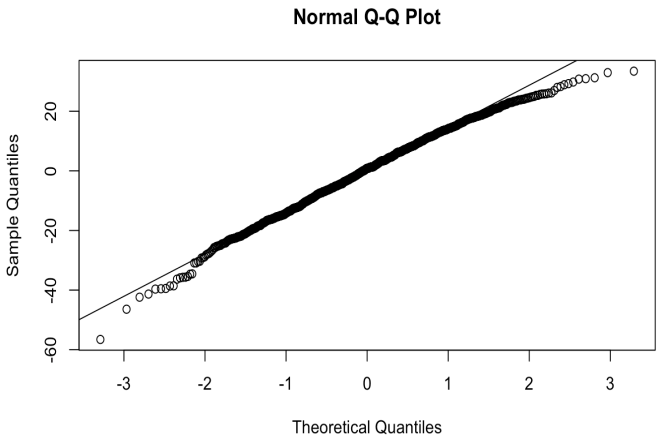
Female students with some high school parents have a largest range between the highest average score and lowest average score followed by female students with master's degree parents, male students with some college parents,female students with Bachelor's degree parents, female students with high school parents, male students with associate's degree parents, female students with associate's degree parents, male students with some high school parents, male students with bachelor's degree parents, male students with high school parents, female students with some college parents, male students with master's degree. As a result, Female students with some high school parents have the highest variability.

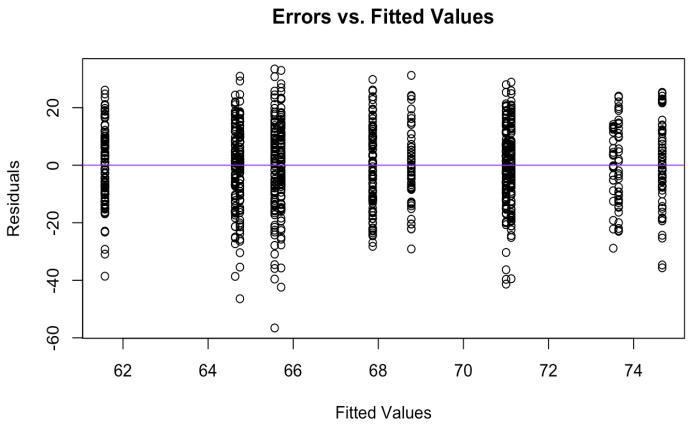
For the differences in median between female and male for each degree, Bachelor's Degree have the largest differences, followed by Some College, Associate's Degree,High School, Some High School and Master's Degree.

Female students with associate's degree parent, female and male students with bachelor's degree,high school,some college and some high school parents, male students with master degree have the outlier. However,the outliers contain less than 5% of the sample size, we do not want to remove that when calculating mean, standard deviation, because this will not significantly impact the values and we don't want to remove every important data.

III-Diagnostics

1) Plots & Test





All of the test-statistics and CIs we make rely on our assumptions that:

(i) All Yij were randomly sampled (Independent).

(ii) All groups are independent.

(iii) $\epsilon ij \sim N(0, stdev = \sigma\epsilon)$ for all i, j.

1-1) P-value table

Shapiro-Wilk normality test	4.728e-05
Brown-Forsythe Test	0.6082

Normality

In the Normal Q-Q Plot, it can be seen that most of the data points are along the diagonal, but at the left and right tails, the data points are significantly off the diagonal, which means that the residuals do not satisfy normality at the tails. This deviation may be due to the influence of outliers. To further verify normality, we conducted the Shapiro Wilk normality test with a p-value of 4.728e-05, which is significantly lower than the significance level (0.05), indicating that the residuals do not conform to the assumption of normal distribution.

Equal Variance

From the Errors vs. Fitted Values graph, it can be seen that most of the residuals are randomly distributed near the zero line, and the vertical dispersion of the residuals is relatively consistent between different fitted values. The residual variance of the model remains basically consistent, meeting the requirements of homoscedasticity. The p-value of the Brown Forsythe test is 0.6082, much greater than 0.05, indicating that we cannot reject the null hypothesis (equal variance), which suggests that the residuals of the model conform to the assumption of Equal Variance.

Conclusion

The normality assumption is violated, but the equal variance assumption is satisfied based on the Brown-Forsythe test so the ANOVA analysis can proceed.

IV-Analysis & Interpretation

1) Model Fit

In this analysis, we are going to state two models to test which model best fits the dataset. The full model and the reduced model. Their existence is to compare the degree of influence of different factors on student performance.

Full Model:

$$Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \epsilon_{ijk}$$

Where:

- μ: Overall mean (Average score)
- **γ**: (Gender): Effect of Gender.
- δ**: (**Parental Level): Effect of Parental Level.
- **γ**δ**:** (Gender×Parental Level): Interaction effect.
- ϵ: Residual error

Reduced Model:

$$Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + \epsilon_{ijk}$$

Where:

- μ: Overall mean (Average score)
- **γ**(Gender): Effect of Gender.
- δ((Parental Level): Effect of Parental Level.
- ϵ: Residual error

Our goal is to determine whether or not Students' Performance are affected by the gender and/or their parent's level of education. The full model includes gender, parental education level, and the interaction effect between gender and parental education level. This means that we not only considered the independent effects of gender and parental education level on student performance, but also the possible interactions between these two factors. For example, will the impact of parental education level on student performance vary depending on the gender of the student. The Reduced Model only considers the independent effects of gender and parental education level on student performance, but ignores their interactive effects. It assumes that gender and parental education level have independent effects on student performance, and there is no interaction between the two.

By comparing these two models, we can determine whether the interaction effect is significant. If the full model performs better than the simplified model, it indicates a significant interaction effect between gender and parental education level. If not, the interaction effect can be ignored. This can help us better understand which factors have a significant impact on student performance and ultimately choose the model we need.

We chose two methods, **interaction plots and F-statistic**, to make the final decision. The interaction plots intuitively show whether there is an interaction between gender and parental education level, and is a preliminary evaluation method to help us determine. The F-statistic provides us with a statistical test, using p-value to determine whether the

interaction effect is statistically significant, thus providing a reliable quantitative basis for our decision-making.

2) Hypothesis Test

To evaluate the effects of Gender and Parental Level on Average Scores, we conducted hypothesis tests for:

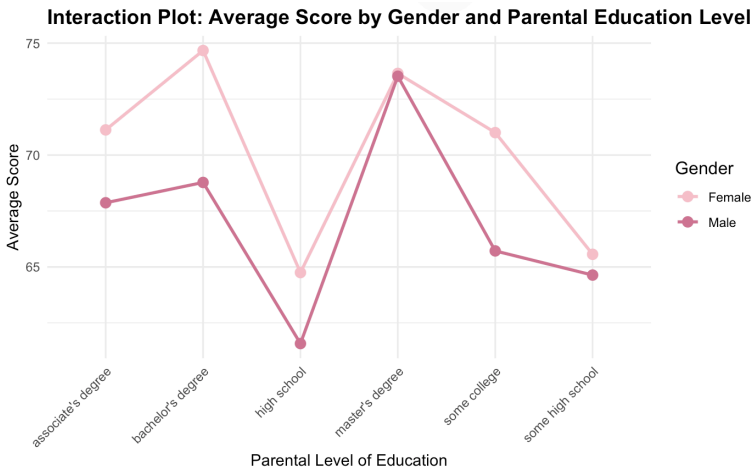
Null Hypothesis (Ho):

- $(\gamma\delta)_{ij} = 0$
- This means that the interaction between Gender (γ\gammaγ) and Parental Level (δ\deltaδ) has no effect on Average Scores. In other words, the effect of one factor (e.g., Gender) does not depend on the levels of the other factor (e.g., Parental Level).

Alternative Hypothesis (Ha):

- at least one $(\gamma\delta)ij \neq 0$
- This means that there is an interaction between Gender and Parental Level, where the effect of one factor depends on the levels of the other.

3) Interaction Plot



The interaction plot above displays the average scores for students based on Gender (Female and Male) and Parental Level of Education. Here's how we interpret it to assess whether there is interaction:

Observations:

Parallel Lines Indicate No Interaction:

- If the lines for Female and Male are mostly parallel, it suggests that the effect of Parental Education on scores is consistent across genders, indicating no significant interaction.
- Here, the lines are relatively parallel, particularly in the lower and higher parental education levels.

Conclusion:

The plot supports the conclusion that there is **no significant interaction** between Gender and Parental Education Level, as the trends across parental education are largely consistent for both genders.

4) ANOVA Table & F-statistic and P-value

	SSE	df	F-Statistic	p-Value
Gender	2904.11	1	15.19	0.0001
Parental Level of Education	9847.00	5	10.30	Less than 0.0001
Gender × Parental Level Interaction	811.15	5	0.85	0.5155
Gender + Parental Level of Education	202494.71	6	11.11	Less than 0.0001

This is the formula for us to form F-Statistic:

$$F_s = \frac{\frac{SSE_R - SSE_F}{df(SSER - df(SSE_F)}}{MSE_F}$$

The F-statistic is used to compare the fitting performance between the full model and the simplified model. By comparing the sum of squared errors (SSE) of these two models, we can determine whether the interaction effect significantly improves the model's fit to the data. Among them, SSE_S represents the sum of squared errors of the simplified model,

SSE-F represents the sum of squared errors of the entire model, df represents the degrees of freedom, and MSE-F represents the mean square error of the entire model. The degrees of freedom for the full SSE is the total number of observations minus sample size of a times b. The degrees of freedom for the reduced SSE is the total number of observations minus sample size of a and b plus 1.

Based on the ANOVA results:

Gender

- **P-value:** 0.0001
- **Decision:** Since the p-value is less than 0.05, we can reject the null hypothesis (gender has no effect on average grades)
- **Decision:** Indicating that gender has a significant impact on students' average grades

Parental Level of Education

- **P-value**: less than 0.0001
- **Decision**: Since the p-value is less than 0.05, we can reject the null hypothesis (Parental Level of Education has no effect on average grades)
- **Decision:**This p-value indicating that Parental Level of Education has a significant impact on students' average grades

Gender × Parental Level of Education (Interaction Effect):

- **P-value**: 0.5155
- **Decision**: Fail to reject the null hypothesis since the p-value is greater than 0.05.
- **Conclusion**: There is no statistically significant interaction effect between gender and parental level of education on average scores. The effect of one factor does not depend on the levels of the other.

Through f-test, we concluded that the interaction effect was not significant, so we chose to use the reduced model. Retaining interaction terms will only increase the complexity of the model without significantly improving the model fitting. Therefore, we have decided to remove interaction effects to make the model more efficient. Reduced model can better reflect the independent effects of gender and parental education level, avoiding complex interaction interference, which helps us to clearly understand the independent impact of each factor.

5) Confidence interval

We chose to analyze pairwise comparisons and contrasts Confidence intervals to gain a deeper understanding of the specific differences in each factor and the significance of differences between different groups. We will focus on each main effect to understand the independent impact of gender and parental education level on student performance. We use the CI formula at a significant level of 0.05.

5-1) Confidence Intervals table

$(1-\alpha)100\%$ for $\mu_{ij} - \mu_{i'j'}$ is

$$\bar{Y}_{ij.} - \bar{Y}_{i'j'.} \pm t_{\alpha/2,df(SSE)}\sqrt{MSE(1/n_{ij} + 1/n_{i'j'})}$$

Pairwise	Difference	Lower bound	Upper bound
F high school – F bachelor's degree	-9.92	-17.29	-2.54
M high school – M bachelor's degree	-7.20	-14.77	0.37
F Master ‘s degree – F highschool	8.89	0.02	17.77
M Master’s degree – M highschool	11.94	1.48	22.40

$(1-\alpha)100\%$ CI for $\sum_j \mu_{.j}$ is

$$\sum_j c_j \hat{\mu}_{j.} \pm t_{\alpha/2,df(SSE)}\sqrt{MSE/a^2 \sum_j c_j^2 (\sum_j \frac{1}{n_{ij}})}$$

Pairwise

F high school – F bachelor's degree

The upper and lower bounds of this confidence interval are both negative, indicating a significant difference between the two groups of female parents with high school and bachelor's degrees. The average score of students from parents with a bachelor's degree is significantly higher than that of students from parents with a high school degree.

M high school – M bachelor's degree

The confidence interval crosses zero, which means we cannot conclude that there is a significant difference between these two groups. However, the average difference is negative (-7.20), indicating that male students from parents with a bachelor's degree may have higher scores, but this difference is not statistically significant as the interval contains zero.

F Master ‘s degree – F highschool & M Master’s degree – M highschool

The upper and lower bounds of the confidence interval are both positive, indicating a significant difference between the two groups of female parents with high school and master's degrees. The average score of students from parents with a master's degree is significantly higher than that of students from parents with a high school degree.

Contrast 1

$Y0,associate + Y0,some\ high\ school - Y1,associate - Y1,some\ high\ school$

This contrasts combined average performance of female students whose parents have either an associate degree or some high school education with the combined average performance of male students under the same parental education level. From the data analysis, we find out the CI is between -5.008547 and 11.51982. Since the Ci includes 0, it indicates that the difference between these groups is not statistically significant. Therefore, there is no significant difference between girls and boys with parental education levels.

Contrast 2

Y0,bachelor+Y1,bachelor+Y0,high school+Y1,high school−Y0,associate−Y1,associate−Y0,some high school

This contrast comparing different combinations of students based on their gender and their parental education level, and from our assessment we have CI between 135.455595 and 169.02188. This CI does not contain 0, indicating that the average performance of the high education group is significantly higher than that of the low education group, which means higher parental education has an impact on student's performance.

V-Conclusion

In this study, we analyzed the effects of gender and parental education level on student performance using a two-way ANOVA. We found that both gender and parental education level independently have significant effects on average student scores. Female students generally outperformed male students, and students with parents having higher levels of education tended to achieve better results. There was no significant interaction effect between gender and parental education level. The diagnostics indicated that while the equal variance assumption for the ANOVA model was satisfied, the normality assumption was not, which might affect the result of our study. Finally,We focus on analyzing the main effects through pairwise comparisons and contrasts to understand more specific differences between groups.

Appendix (Topic 1)

```
`Hawk.(1)`=read.csv("~/Desktop/Hawk (1).csv")
hawks =`Hawk.(1)`
the.model=lm(Wing~Species,data=hawks)
hawks$ei = the.model$residuals

nt = nrow(hawks)
a = length(unique(hawks$Species))
SSE = sum(hawks$ei^2)
MSE = SSE/(nt-a)
eij.star = the.model$residuals/sqrt(MSE)

alpha = 0.01
t.cutoff= qt(1-alpha, nt-a)
rij = rstandard(the.model)
CO.rij = which(abs(rij) > t.cutoff)

# Removing outliers
outliers = CO.rij
new.data = hawks[-outliers,]
new.model = lm(Wing ~ Species,data = new.data)

print(nt)
new.nt=nrow(new.data)
print(new.nt)
library(EnvStats)
qqnorm(new.model$residuals)
qqline(new.model$residuals)

new.ei=new.model$residuals

plot(new.model$fitted.values, new.model$residuals, main = "Errors vs. Group
Means",xlab = "Group Means",ylab = "Errors",pch = 19)
abline(h = 0,col = "purple")

library(ggplot2)
qplot(Species, ei, data = new.data) + ggtitle("Errors vs. Groups") +
xlab("Groups") + ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
library(car)
boxplot(ei ~ Species, data = new.data)
the.SWtest = shapiro.test(new.ei)
the.SWtest
the.BFtest = leveneTest(n.ei~ Species, data=new.data, center=median)
p.val = the.BFtest[[3]][1]
P.val


#Transformation Variable
library(EnvStats)
boxcox(the.model ,objective.name = "PPCC")
boxcox(the.model ,objective.name = "Shapiro-Wilk")
boxcox(hawks$Wing , objective.name = "Log-Likelihood")

L1 =boxcox(the.model ,objective.name = "PPCC",optimize = TRUE)$lambda
L2 = boxcox(the.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda
```

```
L3 = boxcox(hawks$Wing,objective.name = "Log-Likelihood",optimize =
TRUE)$lambda

#Transform Data
lambda=2
YT = (hawks$Wing^(L1)-1)/lambda
t.data = data.frame(Wing = YT, Species = hawks$Species)
t.model = lm(Wing ~ Species,data = t.data)

qqnorm(t.model$residuals)
qqline(t.model$residuals)

t.ei=t.model$residuals

plot(t.model$fitted.values, t.model$residuals, main = "Errors vs. Group
Means",xlab = "Group Means",ylab = "Errors",pch = 19)
abline(h = 0,col = "purple")
library(ggplot2)
qplot(Species, t.ei, data = t.data) + ggtitle("Errors vs. Groups") +
xlab("Groups") + ylab("Errors") + geom_hline(yintercept = 0,col = "purple")

boxplot(t.ei ~ Species, data = t.data)

the.SWtest = shapiro.test(t.ei)
the.SWtest

library(car)
t.BFtest = leveneTest(t.ei~ Species, data=t.data, center=median)
p.val = t.BFtest[[3]][1]
p.val
#Remove outlier and then Transform Data
lambda=2
YT = (new.data$Wing^(L1)-1)/lambda
t.data = data.frame(Wing = YT, Species = new.data$Species)
t.model = lm(Wing ~ Species,data = t.data)

qqnorm(t.model$residuals)
qqline(t.model$residuals)

t.ei=t.model$residuals

plot(t.model$fitted.values, t.model$residuals, main = "Errors vs. Group
Means",xlab = "Group Means",ylab = "Errors",pch = 19)
abline(h = 0,col = "purple")
library(ggplot2)
qplot(Species, t.ei, data = t.data) + ggtitle("Errors vs. Groups") +
xlab("Groups") + ylab("Errors") + geom_hline(yintercept = 0,col = "purple")

boxplot(t.ei ~ Species, data = t.data)
the.SWtest = shapiro.test(t.ei)
the.SWtest
library(car)
t.BFtest = leveneTest(t.ei~ Species, data=t.data, center=median)
p.val = t.BFtest[[3]][1]
p.val
```

```
# Model fit
hawk_data = read.csv("~/Downloads/Hawk.csv")
str(hawk_data)
hawk_data$Species = as.factor(hawk_data$Species)
model <- lm(Wing ~ Species, data = hawk_data)
summary(model)
par(mfrow = c(1, 3))

# Q-Q Plot
qqnorm(residuals(model), main = "Normal Q-Q Plot")
qqline(residuals(model), col = "pink", lwd = 2)

# Residuals vs Fitted
plot(model$fitted.values, residuals(model),
     main = "Residuals vs. Fitted",
     xlab = "Fitted Values",
     ylab = "Residuals",
     pch = 19)
abline(h = 0, col = "pink", lty = 2)
```

Appendix (Topic 2)

```
Mean:
aggregate(average_score~gender+parental_level_of_education,
+ data=Cleaned_Students_Performance,
+ FUN=mean
+ )
Standard deviation:
> aggregate(average_score~gender+parental_level_of_education,
+ data=Cleaned_Students_Performance,
+ FUN=sd
+ )
Sample size:
aggregate(average_score~gender+parental_level_of_education,
+ data=Cleaned_Students_Performance,
+ FUN=length
+ )
Boxplot:
the.means=aggregate(average_score~gender+parental_level_of_education,
data=Cleaned_Students_Performance,mean)
>
the.sds=aggregate(average_score~gender+parental_level_of_education,da
ta=Cleaned_Students_Performance,sd)
>
boxplot(average_score~gender+parental_level_of_education,data=Cleaned
_Students_Performance)
```

```
Cleaned_Students_Performance <-
read.csv("~/Desktop/Cleaned_Students_Performance.csv")
data=Cleaned_Students_Performance

full_model=aov(average_score~gender*parental_level_of_education,data=
data)
# Load necessary libraries
library(ggplot2)
library(car)
students_data = Cleaned_Students_Performance

# Fit the ANOVA model
anova_model = aov(average_score ~ gender +
parental_level_of_education + gender:parental_level_of_education,
data = students_data)
# QQ Plot
qqnorm(residuals(anova_model))
qqline(residuals(anova_model))
# Plotting errors vs fitted values to assess homoscedasticity
fitted_values = fitted(anova_model)
plot(fitted_values, ei, main = "Errors vs. Fitted Values", xlab =
"Fitted Values", ylab = "Residuals")
abline(h = 0, col = "purple")
# Shapiro-Wilk normality test
ei = residuals(anova_model)
shapiro_test = shapiro.test(ei)
print(shapiro_test)
# Brown-Forsythe Test for Homoscedasticity
bf_test = leveneTest(ei ~ parental_level_of_education, data =
students_data, center = median)
print(bf_test)

# Contrast and pairwise
library(multcomp)
library(MASS)
library(DescTools)
data <-  read.csv("~/Downloads/Cleaned_Students_Performance.csv")
data_new <- na.omit(data.frame(Gender = data$gender,
Parental_level = data$parental_level_of_education,Average_score =
data$average_score))
data_new$Gender = as.factor(data_new$Gender)
data_new$Parental_level = as.factor(data_new$Parental_level)
fit.anova = aov(Average_score ~ Gender * Parental_level, data =
data_new)
- TukeyHSD(fit.anova)
```

```
contrasts = list(
  "Male_HighSchool_vs_Female_HighSchool" = c(1, -1, 0, 0, 0, 0, 0, 0,
0, 0),
  "HighSchool_or_Below_vs_College_or_Above" = c(-1, -1, 1, 1, 1, 1,
0, 0, 0, 0),
  "Male_vs_Female_Bachelor" = c(0, 0, 0, 0, 1, -1, 0, 0, 0, 0),
  "Bachelor_vs_Other" = c(-1, 1, 1, 1, 0, 0, 1, 1, 1, 1)
)
scheffe_results
tukey_results = TukeyHSD(fit.anova, "Gender:Parental_level")


# SSE
data = read.csv("~/Downloads/Cleaned_Students_Performance.csv")
data_anova = data.frame(
  Gender = as.factor(data$gender),
  Parental_Level = as.factor(data$parental_level_of_education),
  Average_Score = data$average_score
)
full_model = aov(Average_Score ~ Gender * Parental_Level, data =
data_anova)
reduced_model = aov(Average_Score ~ Gender + Parental_Level, data =
data_anova)
anova_full = summary(full_model)
anova_reduced = summary(reduced_model)
sum_sq_gender = anova_full[[1]]["Gender", "Sum Sq"]
sum_sq_education = anova_full[[1]]["Parental_Level", "Sum Sq"]
sum_sq_interaction <- anova_full[[1]]["Gender:Parental_Level", "Sum
Sq"]
effect_summary <- data.frame(
  Effect = c("Factor A (Gender)", "Factor B (Parental Education)",
"Interaction (A × B)"),
  Sum_of_Squares = c(sum_sq_gender, sum_sq_education,
sum_sq_interaction)
)

# Interaction plot
library(ggplot2)
library(dplyr)

data <- read.csv("~/Downloads/Cleaned_Students_Performance.csv")

data$gender <- factor(data$gender, levels = c(0, 1), labels =
c("Female", "Male"))
```

```r
data$parental_level_of_education <-
as.factor(data$parental_level_of_education)

group_means <- data %>%
  group_by(gender, parental_level_of_education) %>%
  summarize(Average_Score = mean(average_score), .groups = "drop")

ggplot(group_means, aes(x = parental_level_of_education, y =
Average_Score, group = gender, color = gender)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(
    title = "Interaction Plot: Average Score by Gender and Parental
Education Level",
    x = "Parental Level of Education",
    y = "Average Score",
    color = "Gender"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(size = 14, face = "bold"),
    legend.title = element_text(size = 12)
  ) +
  scale_color_manual(values = c("pink", "palevioletred"))
```