Learning Gap during COVID-19 Pandemic

The University of Toronto Mississauga
STA304H5 Lec 0101 Fall 2022

Jingwen (Steven) Shi, Yulong Ding, Zheng Yang Fei,
Jing Mo, Junli Song, Hongsheng Zhong, Xuankui Zhu

Nov 16, 2022

Table of Contents

Abstract	2
Introduction & Research Methodology	3
Variables & Code Book	4
Statistical Analysis	5
Conclusion, Limitation and Future Direction	20
Acknowledgment	21
Appendix	22

1 Abstract

Through the collection and analysis of the data collected, the results showed that there is no significant difference in the quality of online versus in-person learning. Specifically, the average frequency of skipping online and in-person lectures are similar. The p-value of hypothesis testing is high, suggesting that there is no evidence to reject the hypothesis that there is no relation between the two. Other aspects such as the cGPA of students during and after the pandemic and frequency of attending office hours show similar.

2 Introduction & Research Methodology

The coronavirus disease (COVID-19) led to lockdowns and self-isolating practices in Canada and other countries. Universities and schools across the world have changed their course delivery mode from in-person to online learning.

The purpose of this study is to investigate multiple potential factors which may lead to a learning gap during the COVID-19 pandemic, that is, a difference between what students are expected to learn and what students actually learned.

This report used a series of statistical methods to investigate the above relationship from the most basic to advanced methodology in the following sections including but not limited to the confidence interval, hypothesis test, linear, logistic, and univariate regression analysis.

The data being used in this study is manually collected from the University of Toronto Mississauga with a target population of all STA304H5 Fall 2022 students. The sample was collected through an online questionnaire by posting on an internal website for different lecture sections used by the University of Toronto Mississauga. The questionnaires are randomly distributed during various lecture and tutorial sections.

By assuming there are N = 200 students taking STA304H5F in 2022, the sample size is estimated with a margin of error of B = 10% on a 95% confidence interval. Moreover, this study tries to estimate the proportion of students who experienced a learning gap during the COVID-19 will be estimated. Since the data is sampled without replacement by using SRS, this research makes an assumption of having p = q = 0.5.

The computation of the sample size is as follows:

$$n = \frac{Npq}{(N-1)D + pq} = \frac{200 * 1/2 * 1/2}{(200-1)0.1^2/4 + 1/2 * 1/2} = 66.88963 \approx 67$$

However, after removing the undesirable data (e.g. one or more questions being left blank or unreasonable answers provided), there are only 61 valid samples that can be used in this study.

Hence, the margin of error is increased to 11% such that only 58 valid samples are required to draw statistically significant conclusions.

3 Variables and Code Book

Variable	Section	Definition
Pandemic	4.1 - 4.6	The period of time during the COVID-19 Pandemic.
Pre-pandemic	4.1 - 4.6	The period of time before the COVID-19 Pandemic.
Comprehension Lv. of Materials of Online Lecture	4.1	Students' comprehension levels of materials when they take online lectures.
cGPA / CGPA	4.3 - 4.6	Average cumulative GPA of all past school years during the pandemic.
Average Study Time	4.6	Daily average total study time during the pandemic.
Average Frequency of Skipping Lec	4.6	Weekly average frequency of skipping online delivered lectures during the pandemic.
Average Frequency of Attending Office Hours	4.6	Weekly average frequency of attending office hours during the pandemic.
Comprehension Level	4.6	Student's self-recognized/assessed comprehension level of materials of online lecture during the pandemic.
Distraction Lv. of Online Lec Compared to In-Person Lec	4.3	The level of distraction students experience while receiving online instruction.

4 Statistical Analysis

4.1 Basic Graph Analysis

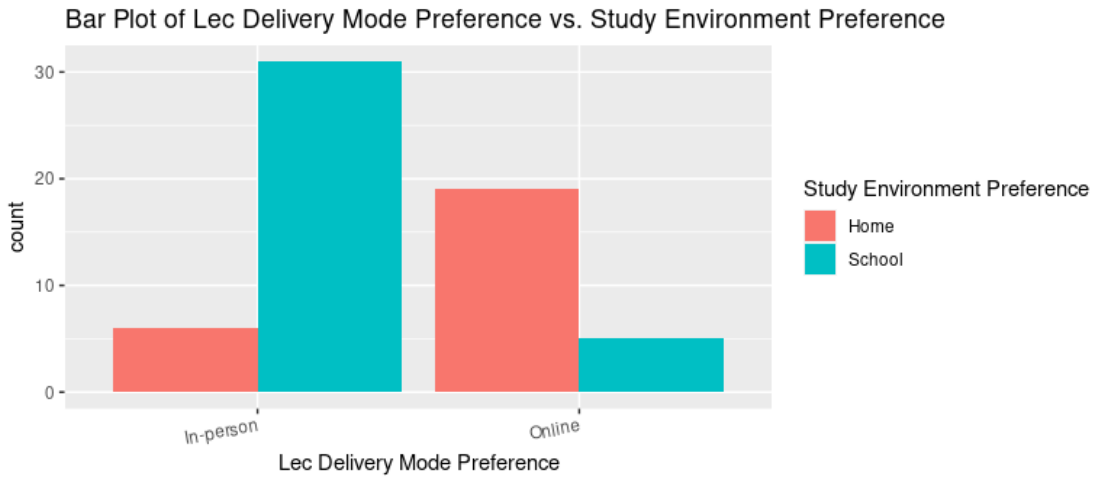


Figure 1
Bar plot of Lec Delivery Mode Preference vs. Study Environment Preference

The bar plot above shows a highly collaborated preference between lecture delivery method and study environment, that is, the majority of students who prefers to study at school also prefer in-person lecture and vice versa.

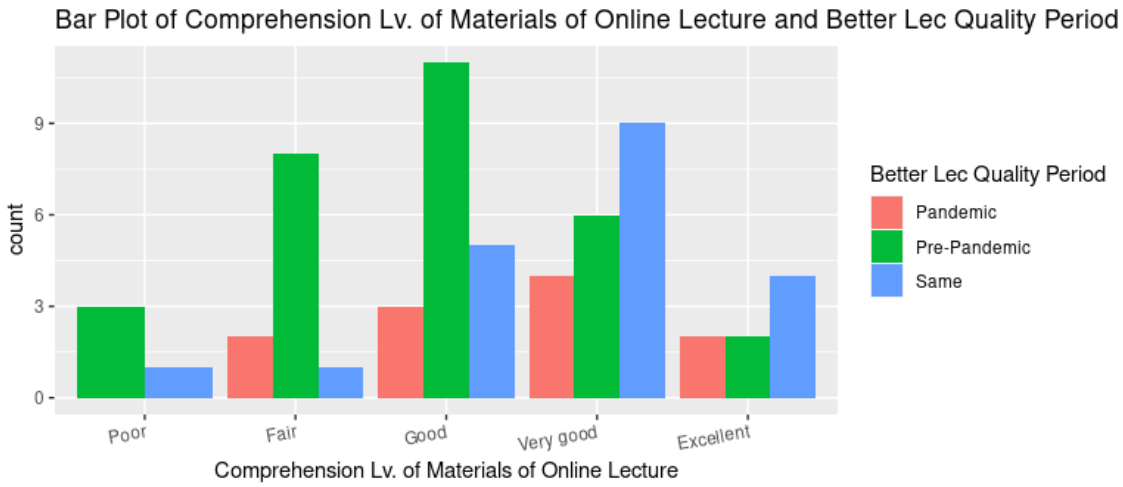


Figure 2
Bar plot of comprehension Lv. of materials of online lecture and better lec quality period

In Figure 2, as students have a higher level of comprehension in the course materials, students tend to believe the lecture qualities are the same and vice versa. Those students who understand the course material better might have a high concentration or earnest steadfast study manner. However, those explanations need more data to support this hypothesis.

4.2 Linear Regression Analysis

In this section, the linear regression model will be used to investigate the relationship between the average frequency of skipping online lectures and in-person lectures.

If students' preferences for both course delivery methods are the same, then there should be an approximately one-to-one ratio between these two variables. This can be verified and compared with the proportion of students who believe in-person lecture's quality is better than that of online lectures and vice versa.

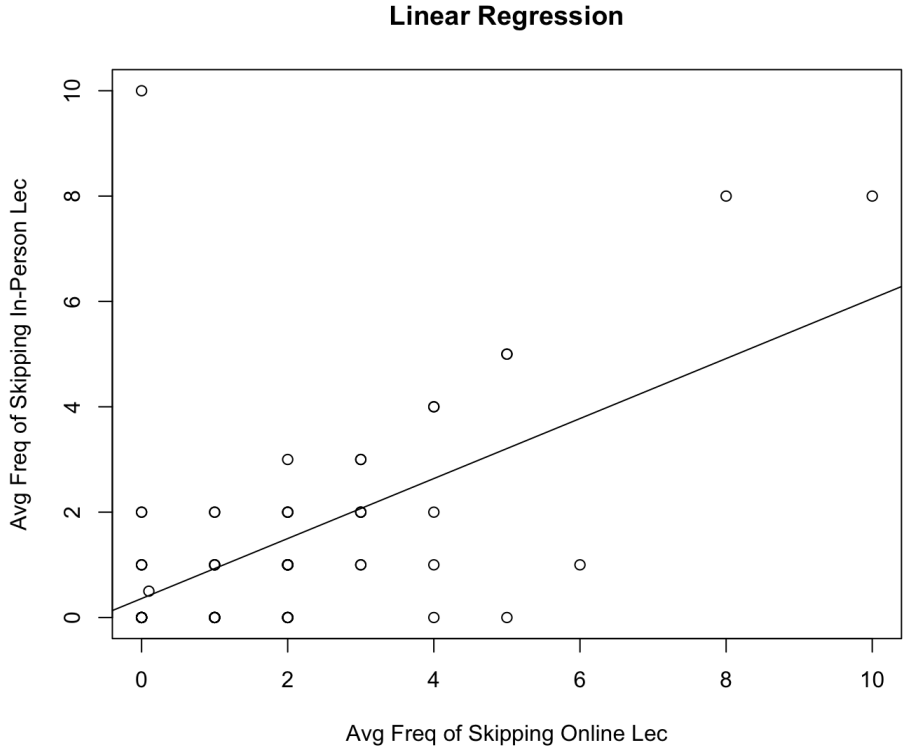
4.2.1 Beta Estimates

Let X = Average frequency of skipping online lectures.
Y = Average frequency of skipping in-person lectures.

Then the unbiased estimators of linear regression are as follows:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = 0.5696$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 0.3599$$



4.2.2 Hypothesis Testing of Coefficient & ANOVA Table

In order to test if there is a linear correlation between the two variables, the ANOVA table will be useful to illustrate.

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_a : \beta_1 \neq 0$$

ANOVA Table

Source	DF	Sum of Square	Mean Square	F value	Pr(>F)
Regression	1	SSR = 79.687	MSR = SSR = 79.687	MSR / MSE = 26.075	3.695e-06
Error	59	SSE = 180.313	MSE = SSE/(n-2) = 3.056		

Since the p-value obtained by using an F-distribution test is less than 0.05, H_0 will be rejected. Hence, there is strong evidence to show a correlation between the two variables.

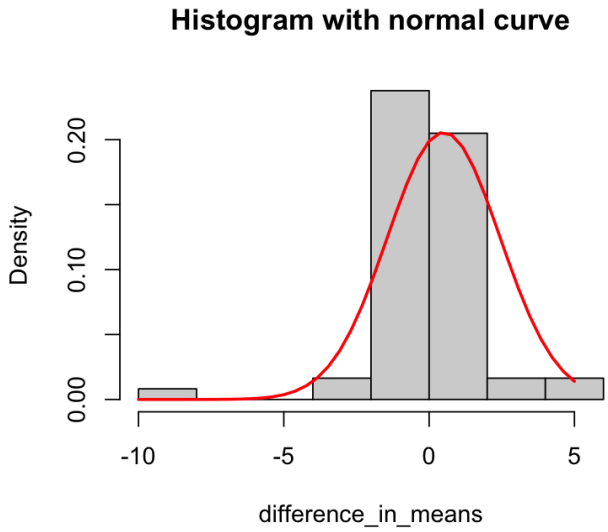
4.2.3 Hypothesis Testing of Mean

In an effort to determine whether COVID-19 would cause a learning gap, the hypothesis test on the average frequency of skipping online lectures and the average frequency of skipping in-person lectures is constructed.

Construct hypothesis tests and compare two average frequencies to check whether there is a relationship between two variables.

Sample 1 is the average frequency of skipping online lectures weekly and Sample 2 is the average of skipping in-person lectures weekly. Observations in sample 1 are highly correlated with observations in sample 2, so the data are matched pairs. For each pair, the difference in means is $\mu_1 - \mu_2$.

Assuming the sample of differences follows a Normal model, the following histogram is constructed. There is a clear normal curve. Hence, the sample of difference follows a normal distribution.



Assuming there is no learning gap during COVID-19, the null hypothesis is that the average frequencies of skipping online lectures and in-person lectures are similar. This study considers the following hypothesis:

$$H_0: \mu_1 = \mu_2 \qquad VS \qquad H_1: \mu_1 \neq \mu_2$$

Use $\alpha = 0.05$ as a significance level.

Paired T-Test

Variables	t-statistics	df	p-value	LCL	UCL
avg_freq_skip_online	2.0196	60	0.0479	0.0048	0.9985
avg_freq_skip_inperson					

Since the p-value equals 0.0479 and it is smaller than the alpha level, this leads to rejecting the null hypothesis.

Therefore, it is not sufficient to conclude that the average frequencies of skipping online lectures and in-person lectures are similar.

4.2.4 Comparison between Self-Assessed Better Lecture Delivered Method and Linear Model

In the figure below, it is obvious that the number of students who think the pandemic period has a better teaching quality is around a third of the ones who think pre-pandemic period has a better teaching quality.

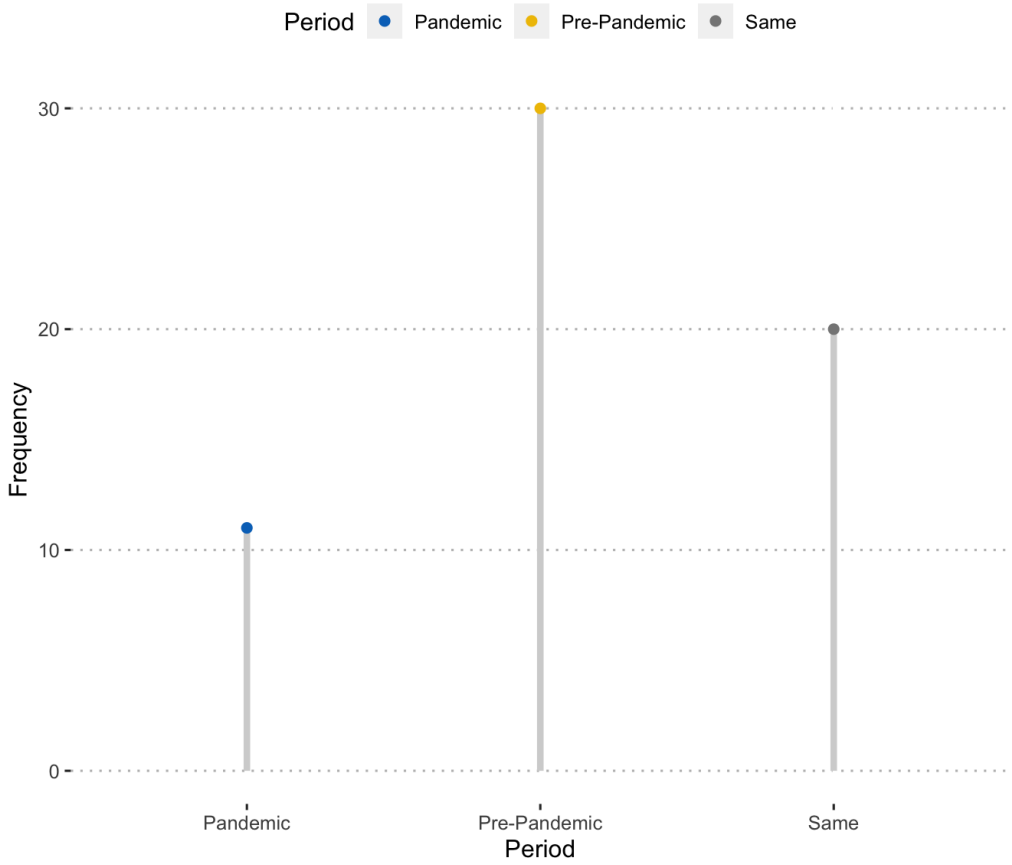


Figure 1: Dot Plot of Period with Better Quality Lectures

This observation might be one of the reasons that there are more students who have a higher frequency of skipping online delivered lectures than in-person lectures.

4.2.5 Conclusion

Through the model we can find that the number of people who skip classes offline is not the same as the number of people who skip classes online. It is possible that the number of people who skip classes offline is less than that of online because they are online with playback, or it is possible that the quality of online courses is not as good as that of in-person, and it is not as good as spending time on self-study so they choose to skip classes, which needs further study.

4.3 Logistic Regression Analysis - After the Pandemic

This section will examine the relationship between students' cGPA and multiple factors during the pandemic using logistic regression.

4.3.1 Conditions

1. By definition, the response variable cGPA is a categorical variable of an ordinal number. Among the three independent variables, the degree of distraction in class is categorical. And the average frequency of attending office hours and frequency of skipping classes are continuous variables.

2. Determining whether a regression model can be used requires the use of the Variance Inflation Factor to measure the correlation and strength of correlation between predictor variables in the regression model.

Variance Inflation Factor Test

	GVIFA	DF	$GVIF^{1/(2*Df)}$
avg_freq_skip	1.025790	1	1.012813
avg_freq_OH	1.039917	1	1.019763
comp_lv	1.038637	2	1.009522

For the coefficients of all variables, the variance inflation factor was below 5 and around 1. Therefore, the data is strong enough to conclude that there is no multicollinearity between the independent variables.

4.3.2 Brand test to determine if mode follows parallel regression assumption.

Brand Test

Test for	X2	DF	Probability
Omnibus	3.14	4	0.32
comp_lvLow	2.13	2	0.3
comp_lvMid	0.49	2	0.78

Since the probability and Omnibus of both comp_lvLow and comp_lvMid are greater than 0.05. Therefore, the parallel regression assumption holds.

4.3.3 Univariate 95% confidence interval

Modeling only one independent variable at a time with the response variable, the coefficients of each variable, and their corresponding 95% confidence intervals can be obtained.

Variable	LCL	Coefficient	UCL
avg_freq_skip	–0.1857877	0.06884668	0.3238837
avg_freq_OH	–0.6672023	–0.2531661	0.1378202
comp_lvLow	–2.868074	–0.2531661	0.3019740
comp_lvMid	–1.471055	0.1033543	0.9088668

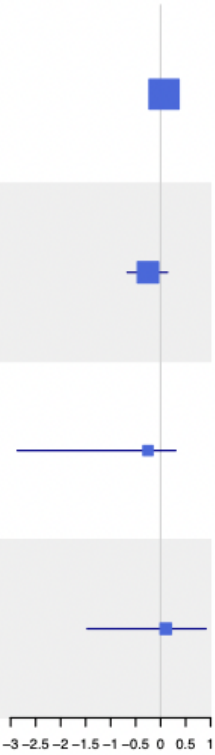


Figure 4.3: Univariate 95% confidence interval,

Unfortunately, Figure 4.3 shows that all variables have a 95 percent confidence space containing zeros, which means that all variables should no longer be considered for the rest of the test.

In order to further explore the effect of retaining Distraction Lv. of Online Lec Compared to In-Person, Avg Freq of Skipping In-Person Lec (Weekly), as well as Avg Freq of Participating OH (After Pandemic), are removed. Person Lec (Low) and Distraction Lv. of Online Lec Compared to In-Person Lec (Mid).

4.3.4 Coefficients Analysis

Coefficients	
Variable	Coefficient
comp_lvLow	0.3766
comp_lvMid	1.0517

The coefficients are expressed, with all other variables in the model held constant.

- For the degree of distraction for Low, a decrease of 1.2523069 is expected on the log scale.
- For the degree of distraction for Mid, a decrease of 0.2618523 is expected on the log scale.

4.3.5 Conclusion

As the data shows that the comp_lvMid has a strong association with cGPA, but since the confidence interval of all variables contains 0, the model may have a high error. Further investigation is needed for the factors that really have an impact on cGPA.

4.4 Two-Variable Analysis

To investigate the academic impact of the pandemic, differences in the proportions of students with high cGPA during and after the pandemic are analyzed. For convenience, cGPA scores of 3.0 or higher are considered high cGPAs, while those that are below are considered as low cGPAs.

4.4.1 Graphical Analysis

The frequencies of high and low cGPA scores during and after the pandemic are shown in Figure 5.1 and Figure 5.2 respectively.

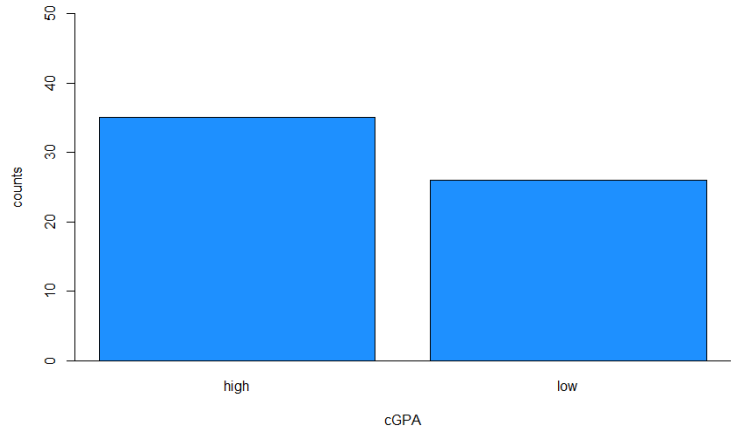


Figure 5.1: Frequencies of high and low cGPA during the pandemic

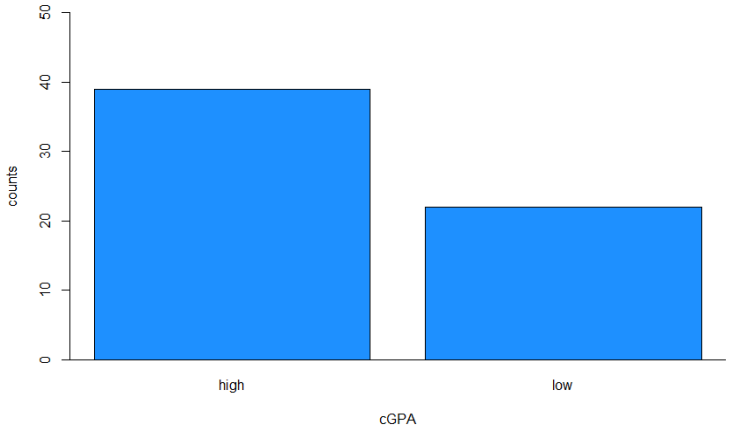


Figure 5.2: Frequencies of high and low cGPA after the pandemic

4.4.2 Hypothesis Testing

Since the cGPA after the pandemic is likely to reflect cGPA during the pandemic, the two proportions are certainly not independent from each other. Therefore, a McNemar's test is conducted.

		After Pandemic	
		High	Low
During Pandemic	High	29(a)	5(b)
	Low	10(c)	15(d)

The frequencies of student records with specific cGPA (high or low) during and after the pandemic are recorded in the above 2 by 2 table. For example, there are 29 students who maintained high cGPA both during and after the pandemic. Each cell is denoted by a letter from a to d. We are interested in finding whether there is a significant difference between the proportion of students who dropped from high to low cGPA(b), and the proportion of students who increased their cGPA from low to high(c)

H0: pb = pc vs H1: pb ≠ pc

α = 0.05

Test Statistic:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Test result:

chi-squared	df	p-value
1.6667	1	0.1967

Under α = 0.05, the p-value is greater than α. Therefore, we accept H0, and conclude that there is no significant change in cGPA during and after the pandemic.

4.5 Ordinal Logistic Regression - During Pandemic

With all the basic analyses and tests being made on the estimator/estimates in the previous sections, this part will investigate the relationship between a student's cGPA and multiple factors during the pandemic (refer to the code book on section 4.6) by using logistic regression.

4.5.1 Assumptions & Conditions

1. The response variable is categorical and measured on an ordinal level.

   In this section, cGPA is the response variable, which is ordinal by its definition. Therefore, the assumption holds.

2. Exist one or more independent variables which are either continuous, categorical or ordinal.

   There are three independent variables in this section where Average Study Time is a categorical variable and Average Frequency of Attending Office Hours and Average Frequency of Skipping Lec are continuous variables. Hence, the assumption holds true.

3. No Multicollinearity between independent variables.

Variance Inflation Factor Test

	GVIF	DF	$GVIF^{1/(2*Df)}$
avg_study_time	1.062694	3	1.010186
avg_freq_skip	1.029851	1	1.014816
avg_freq_OH	1.036753	1	1.018211

For the coefficients of all variables, the variance inflation factor (VIF) is below ten and around one. Therefore, it is sufficient to conclude that there is no multicollinearity between independent variables.

4. Proportional Odds Assumption / Parallel Regression Assumption

The Brant test is a simple way to assess proportional odds for ordinal logistic regression in R. If the p-value < 0.05, then reject the H_0: Parallel Regression Assumption holds.

Brant Test

Test For	X2	df	Probability
Omnibus	7.46	6	0.28
avg_study_time: 3 - 4 Hours	2.86	2	0.24
avg_study_time: 4 - 5 Hours	0	2	1
avg_study_time: > 5 Hours	0.48	2	0.79

Since all variables' p-values are greater than 0.05, it is reasonable to assume that this ordinal logistic model follows the parallel regression assumption hold.

Note: The proportional odds assumption is verified after the modification of the model. Please refer to section 4.6.2.

4.5.2 Univariate Regression Analysis & Variable Screening

Variable	LCL	Coefficient	UCL
avg_study_time: 3 - 4 Hours	-0.8527867	0.3766106	1.6107616
avg_study_time: 4 - 5 Hours	-0.1937557	1.0517438	2.3431600
avg_study_time: > 5 Hours	-3.6680004	-2.0211190	-0.5102267
avg_freq_skip	-0.1368487	0.1033543	0.3374198
avg_freq_OH	0.2317010	0.2040784	0.6477571

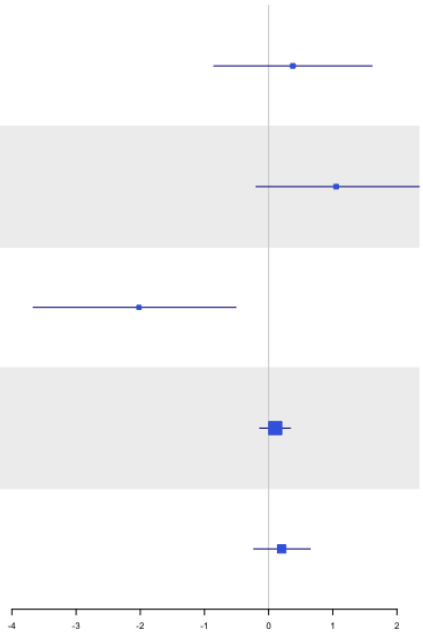


Figure 1: 95% Confidence Interval of Coefficients of Univariate Regressions

By modeling only one independent variable with the response variable at a time (i.e. univariate modeling), the coefficient of each variable and their corresponding 95% confidence interval is shown in the above graph (Figure 1).

There are four variables' confidence intervals of the coefficient including zero, that is, the average frequency of skipping online lectures, the average frequency of attending office hours, average study time of 3 - 4 and 4 - 5 hours.

Therefore there is no statistical significance in the univariate models with the variables mentioned above and will not be considered in the model anymore. However, the average study time of 3 - 4 and 4 - 5 hours will be kept for future analysis and comparisons.

4.5.3 Coefficients & Odds Ratio

Coefficients and OR Summary Table

Variable	Coefficient	OR	LCL	UCL
avg_study_time: 3 - 4 Hours	0.3766	1.4573368	0.42622552	5.0066228
avg_study_time: 4 - 5 Hours	1.0517	2.8626387	0.82385915	10.4140929
avg_study_time: > 5 Hours	-2.0211	0.1325071	0.02552746	0.6003595

The 95% confidence intervals include no 0 for all coefficients, therefore, this new model after variable screening does have statistical significance.

Moving forward, the coefficients represent that, given all other variables remain constant in the model:

- For an average study time of 3 - 4 hours, it is expected to have a 0.38 increase on the log scale.
- For an average study time of 4 - 5 hours, it is expected to have a 1.05 increase on the log scale.
- For an average study time of more than 5 hours, it is expected to have a 2.02 decrease on the log scale.

The odds ratio provides a quantitative measurement of associations between two events. It is the ratio between given an event A and missing B, and given an event B and missing A.

If the OR value equals 1, then there is no association between the two events such that they are independent. Otherwise, if the OR value is greater than 1, then two events are positively correlated and the presence of event A will increase the odds of B and vice versa. Self-evidently, the case of OR value less than -1 is opposite to the previous case.

For all other variables holding constant, the odds ratio shows that:

- For students who study 3 to 4 hours a day, the odds of being more likely to increase in cGPA is 1.146 times more than the ones who do not.
- For students who study 4 to 5 hours a day, the odds of being more likely to increase in cGPA is 2.86 times more than the ones who don't.
- For students who study 5 to 6 hours a day, the odds of being more likely to increase in cGPA is 86.75% (i.e. $(1 - 0.1325071) \times 100\%$) times lower than the ones who do not.
    - For students who do not study 5 to 6 hours a day, the odds of being more likely to increase in cGPA is 7.55 (i.e. 1 / 0.1325071) times more than the ones who do.

Therefore, the statistic above shows that study time does not necessarily have a positive correlation with one's cGPA. It seems that 5 hours per day is a threshold and there might be other factors/reasons for this unexpected observation that need further investigation.

Conclusion, Limitation and Future Direction

In conclusion, the average frequencies of skipping online and in-person lectures are similar, so there might be no learning gap during the pandemic. Furthermore, cGPA is quantitative data that can directly show if a student understands knowledge.

For the analysis of ordinal logistic regression during the pandemic, the statistic shows that the study time does not necessarily have a positive correlation with one's cGPA and there seems no other factors will affect one's cGPA. For the logistic regression after the pandemic, the data shows that the comp_lvMid has a strong association with cGPA. But since the confidence interval of all variables contains 0, the model may have errors. Hence, there is no significant evidence to show that there is learning during the pandemic.

However, due to the limitation of time and of data collection, the error bound can not be reduced furthermore. In the future, it will be interesting to optimize a logistic regression machine learning model to predict one's cGPA by giving the student's study time, participation in lectures, frequency of skipping lectures and so on.

ACKNOWLEDGMENT

We, as group members of Group 2 who are taking STA304H5F in 2022, would like to acknowledge and appreciate all the participants of this statistical analysis. Our sincere thanks to especially the following:

Prof. Luai Al Labadi, Xinyi Yao, Mark Asuncion, and Anna Ly.

We would like to give our warmest thanks to Prof. Luai AI Labadi for his wonderful lectures and detailed explanations of statistical proofs. In addition, we are truly thankful to Xinyi Yao and Mark Asuncion for answering our questions patiently. Moreover, we genuinely appreciate the detailed instructions and feedback which motivated and encouraged us to work hard in this course and in the future.

Appendix

Section 4.2.1, 4.2.2, 4.4.4

```
# Read ME
# Import STA304_Group2_Clean_Dataset.csv from Google Sheet
# Please follow the instructions in the terminal and give full access of your Google account
# If you failed to authorize TidyVerse API, you can reopen RStudio and run the file again and choose 0 in the terminal to generate a new access token
library(readr)
library(googlesheets4)
library(foreign)
library(MASS)
library(Hmisc)
library(reshape2)
library(car)
library(broom)
library(VGAM)
library(forestplot)
library(dplyr)
library(ggplot2)
library(ggpubr)
theme_set(theme_pubr())

path = "https://docs.google.com/spreadsheets/d/1mSdhVUqRZIgQaqskdE_cbvpxD29HAk2WNod3Rxxu2PQ/edit?usp=sharing"
dataset <- read_sheet(path)

# Define Variables
avg_skip_in_person = dataset$`Avg Freq of Skipping In-Person Lec (Weekly)`
avg_skip_online = dataset$`Avg Freq of Skipping Online Lec (Weekly)`

# Build linear regression model
lr = lm(avg_skip_in_person ~ avg_skip_online)

summary(lr)
anova(lr)
```

```
dataset = read.csv("STA304_Group2_Clean_Dataset.csv")
cgpa.during = dataset$cGPA.during.the.Pandemic..2019.Winter...2021.Winter.
cgpa.after = dataset$cGPA.after.the.Pandemic..2021.Spring...2022.Summer.
during <- data.frame(gpa = cgpa.during)
after <- data.frame(gpa = cgpa.after)
during <- transform(
  during,
  category =
    ifelse(
      gpa %in% c("3.0 - 3.5", "3.6 - 4.0"),
      "high",
      "low"
    )
)

after <- transform(
  after,
  category =
    ifelse(
      gpa %in% c("3.0 - 3.5", "3.6 - 4.0"),
      "high",
      "low"
    )
)
during.category <- during$category
after.category <- after$category

#figure 6
barplot(table(during.category),
        ylim = c(0,50),
        xlab = "cGPA",
        ylab = "counts",
        col = "dodgerblue")
abline(h = 0)

#figure 7
barplot(table(after.category),
        ylim = c(0,50),
        xlab = "cGPA",
        ylab = "counts",
        col = "dodgerblue")
abline(h = 0)
```

```
dataset <- dataset %>% mutate(cGPA =
case_when((cGPA.during.the.Pandemic..2019.Winter...2021.Winter. %in% c("3.0 - 3.5", "3.6 - 4.0"))
                                & (cGPA.after.the.Pandemic..2021.Spring...2022.Summer. %in% c("3.0 -
3.5", "3.6 - 4.0")) ~ "high to high",
                                (cGPA.during.the.Pandemic..2019.Winter...2021.Winter. %in% c("1.0 - 1.5",
"2.0 - 2.5", "2.6 - 2.9"))
                                & (cGPA.after.the.Pandemic..2021.Spring...2022.Summer. %in% c("3.0 -
3.5", "3.6 - 4.0")) ~ "low to high",
                                (cGPA.during.the.Pandemic..2019.Winter...2021.Winter. %in% c("3.0 - 3.5",
"3.6 - 4.0"))
                                & (cGPA.after.the.Pandemic..2021.Spring...2022.Summer. %in% c("1.0 -
1.5", "2.0 - 2.5", "2.6 - 2.9")) ~ "high to low",
                                (cGPA.during.the.Pandemic..2019.Winter...2021.Winter. %in% c("1.0 - 1.5",
"2.0 - 2.5", "2.6 - 2.9"))
                                & (cGPA.after.the.Pandemic..2021.Spring...2022.Summer. %in% c("1.0 -
1.5", "2.0 - 2.5", "2.6 - 2.9")) ~ "low to low"))


frequency.table <- as.data.frame(table(dataset$cGPA))
frequencies <- frequency.table$Freq
two.by.two.table <- matrix(frequencies, nrow = 2,
                    dimnames = list("After Pandemic" = c("High", "Low"),
                                "During Pandemic" = c("High", "Low")))
mcnemar.test(two.by.two.table, correct = TRUE)
```

```
# Re-leveling
dataset$`Distraction Lv. of Online Lec Compared to In-Person Lec`<- factor(dataset$`Distraction Lv. of Online Lec Compared to In-Person Lec`, levels = c('1', '2', '3', '4', '5' ), labels = c("Poor", "Fair", "Good", "Very good", "Excellent"), ordered = TRUE)

dataset$`Comprehension Lv. of Materials of Online Lecture`<-factor(dataset$`Comprehension Lv. of Materials of Online Lecture`, levels = c('1', '2', '3', '4', '5' ), labels = c("Poor", "Fair", "Good", "Very good", "Excellent"), ordered = TRUE)

ggplot(dataset, aes(x = factor(dataset$`Lec Delivery Mode Preference`),
                    fill = factor(dataset$`Study Environment Preference`))) +
  geom_bar(position = "dodge") +
  labs(fill = "Study Environment Preference") +
  xlab("Lec Delivery Mode Preference") +
  ggtitle("Bar Plot of Lec Delivery Mode Preference vs. Study Environment Preference") +
scale_x_discrete(guide = guide_axis(angle = 10))

factor(dataset$`Comprehension Lv. of Materials of Online Lecture`),
                    fill = factor(dataset$`Better Lec Quality Period`))) +
  geom_bar(position = "dodge") +
  labs(fill = "Better Lec Quality Period") +
  xlab("Comprehension Lv. of Materials of Online Lecture") +
  ggtitle("Bar Plot of Comprehension Lv. of Materials of Online Lecture and Better Lec Quality Period")
+ scale_x_discrete(guide = guide_axis(angle = 10))
```

```
# Linear Reg Plot
plot(avg_skip_in_person ~ avg_skip_online, data = dataset, main="Linear Regression", xlab="Avg Freq
of Skipping Online Lec", ylab="Avg Freq of Skipping In-Person Lec")
abline(lr)

table(dataset$`Better Lec Quality Period`)

Period = c("Pandemic", "Pre-Pandemic", "Same")
Frequency = c(11, 30, 20)

df <- data.frame(Period, Frequency)

ggplot(df, aes(Period, Frequency)) +
  geom_linerange(
    aes(x = Period, ymin = 0, ymax = Frequency),
    color = "lightgray", size = 1.5
  )+
  geom_point(aes(color = Period), size = 2)+
  ggpubr::color_palette("jco")+
  theme_pubclean()
```

Section 4.5 & 4.3:

```
# Read ME
# Import STA304_Group2_Clean_Dataset.csv from Google Sheet
# Please follow the instructions in the terminal and give full access of your Google account
# If you failed to authorize TidyVerse API, you can reopen RStudio and run the file again and choose 0 in the terminal to generate a new access token
library(readr)
library(googlesheets4)
library(foreign)
library(ggplot2)
library(MASS)
library(Hmisc)
library(reshape2)
library(car)
library(broom)
library(VGAM)
library(forestplot)
library(dplyr)

path = "https://docs.google.com/spreadsheets/d/1mSdhVUqRZIgQaqskdE_cbvpxD29HAk2WNod3Rxxu2PQ/edit?usp=sharing"
dataset <- read_sheet(path)

# Response Variable
cGPA = dataset$`cGPA during the Pandemic (2019 Winter - 2021 Winter)`

# Predictor Variable
avg_study_time = dataset$`Avg Study Time (During Pandemic)`
avg_freq_skip = dataset$`Avg Freq of Skipping Online Lec (Weekly)`
avg_freq_OH = dataset$`Avg Freq of Participating OH (During Pandemic)`
comp_lv = dataset$`Comprehension Lv. of Materials of Online Lecture`

# Initialize training set
training_set <- data.frame(cGPA, avg_freq_OH, avg_freq_skip, avg_study_time, comp_lv)

training_set$cGPA <- factor(training_set$cGPA, levels = c('3.6 - 4.0', '3.0 - 3.5', '2.6 - 2.9', '2.0 - 2.5', '1.6 - 1.9', '1.0 - 1.5', '0.6 - 0.9', '0.0 - 0.5'), labels = c("Excellent", "Good", "Adeqate", "Poor", "Poor", "Poor", "Poor", "Poor"), ordered = TRUE)
training_set$avg_study_time <- factor(training_set$avg_study_time, levels = c('< 3 Hours', '3 - 4 Hours', '4 - 5 Hours', '5 - 6 Hours', '> 6 Hours'), labels = c(': < 3 Hours', ': 3 - 4 Hours', ': 4 - 5 Hours', ': > 5 Hours', ': > 5 Hours'))

# Check assumption of No Multi-collinearity
```

```
olr <- polr(as.factor(cGPA) ~ avg_study_time + avg_freq_skip + avg_freq_OH, data = training_set,
Hess=TRUE)
vif(olr)

# Single Factor Analysis
study_time_fit = polr(as.factor(cGPA) ~ avg_study_time, data = training_set)
freq_skip_fit = polr(as.factor(cGPA) ~ avg_freq_skip, data = training_set)
freq_OH_fit = polr(as.factor(cGPA) ~ avg_freq_OH, data = training_set)

drop1(study_time_fit,test = "Chi")
drop1(freq_skip_fit,test = "Chi")
drop1(freq_OH_fit,test = "Chi")


# Confidence Interval of Coef
confint(study_time_fit, level = 0.95)
coef(summary(study_time_fit))

confint(freq_skip_fit, level = 0.95)
coef(summary(freq_skip_fit))

confint(freq_OH_fit, level = 0.95)
coef(summary(freq_OH_fit))

base_data <- tibble::tibble(mean  = c(0.3766106, 1.0517438, -2.0211190, 0.1033543, 0.2040784),
                        lower = c(-0.8527867, -0.1937557, -3.6680004, -0.1368487, -0.2317010),
                        upper = c(1.6107616, 2.3431600, -0.5102267, 0.3374198, 0.6477571),
                        coef_name = c("avg_study_time: 3 - 4 Hours", "avg_study_time: 4 - 5 Hours",
"avg_study_time: > 5 Hours", "avg_freq_skip",
                                "avg_freq_OH"),
                        coef  = c('0.3766106', '1.0517438', '-2.0211190', '0.1033543', '0.2040784'),
                        LCL = c('-0.8527867', '-0.1937557', '-3.6680004', '-0.1368487', '0.2317010'),
                        UCL = c('1.6107616', '2.3431600', '-0.5102267', '0.3374198', '0.6477571'))

base_data |>
  forestplot(labeltext = c(coef_name, LCL, coef, UCL),
             xlog = FALSE) |>
  fp_set_style(box = "royalblue",
               line = "darkblue",
               summary = "royalblue") |>
  fp_add_header(coef_name = c("", "Variable"),
                coef = c("", "Coefficient"),
                LCL = c("", "LCL"),
                UCL = c("", "UCL")) |>
```

```
  fp_set_zebra_style("#EFEFEF")

# Proportional Odds Assumption
olr <- polr(as.factor(cGPA) ~ avg_study_time, data = training_set, Hess=TRUE)
brant::brant(olr)

summary(olr)
exp(cbind(OR = coef(olr), confint(olr)))
```

Section 4.2.3

```
#Hypothesis Test
STA304_Group2_Clean_Dataset <- read.csv("~/Desktop/STA304_Group2_Clean_Dataset.csv",
header=TRUE)
online_avg=c(STA304_Group2_Clean_Dataset$Avg.Freq.of.Skipping.Online.Lec..Weekly.)
inperson_avg=c(STA304_Group2_Clean_Dataset$Avg.Freq.of.Skipping.In.Person.Lec..Weekly.)
difference=online_avg-inperson_avg

#Histogram
hist(difference_in_means, prob = TRUE, main = "Histogram with normal curve")
x = seq(min(difference_in_means), max(difference_in_means), length = 40)
f = dnorm(x, mean = mean(difference_in_means), sd = sd(difference_in_means))
lines(x, f, col = "red", lwd = 2)
```