Analyzing the effectiveness of Covid-19 vaccines among different age groups using Multinomial logistic regression Model

A Thesis
Submitted to the Faculty of Graduate Studies and Research
In Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Statistics
University of Regina

by
Arfa Khalid
Regina, Saskatchewan
May 2023

Abstract

This study is conducted to evaluate the effectiveness of Covid-19 vaccines in different age groups in Saskatchewan, Canada. Data was collected between September 2021 and December 2021, and a statistical method called multinomial logistic regression was used to analyze the relationships between multiple categorical variables. In this study, the categorical variables were the age groups and the vaccination status (fully vaccinated cases, partially vaccinated cases, and unvaccinated cases) of the individuals with the interaction effect of rate of cases. The mathematical proof for the multinomial logistic regression model with interaction effect was derived in this study.The study demonstrated the effectiveness of Covid-19 vaccines among vaccinated age groups and provided theory and practical application of the multinomial logistic regression model. Results show that there is a statistically significant impact of age group and vaccination status on the effectiveness of Covid-19 cases in Saskatchewan. Specifically, there is a difference in vaccine effectiveness based on age groups and vaccination status.The findings of this study provide crucial insights for policymakers and public health officials to optimize vaccination rollout strategies and control the spread of Covid-19. Overall, this study represents an important step in the ongoing efforts to understand the effectiveness of Covid-19 vaccines and to develop policies and interventions that can help mitigate the pandemic impact.

Acknowledgements

I would like to extend my sincere appreciation to my supervisor Dr. Dianliang Deng for his invaluable guidance, caring patience and all help and support provided for me over my graduate study. During courses period, Dr. Deng's exquisite professional knowledge has laid a solid theoretical foundation for me. During research phase, Dr. Deng not only gave professional guidance in the research, but also cultivated my ability to explore and solve problems, which will benefit me all the life. In addition, I would also like to express my gratitude to Dr. Taehan Bae, Dr. Michael Kozdron and Dr. Arzu sardarli for the fascinating courses, which enhanced my professional knowledge. In addition, I am indebted to the Department of Mathematics and Statistics and Faculty of Graduate Studies and Research for scholarships and teaching assistant job opportunities.

Dedication

This thesis is dedicated to my family whose unwavering support and boundless love have been my guiding light throughout my journey. Their commitment to responsibility and diligence has instilled in me a passion for academic excellence, and I am forever grateful for their inspiration. Without their encouragement, I could not have accomplished what I have today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The Covid-19 pandemic has affected the world in unprecedented ways, causing significant health, social, and economic impacts. Since the first cases were reported in Saskatchewan in March 2020, the province has been taking measures to mitigate the spread of the virus, including vaccination campaigns. Saskatchewan has approved several Covid-19 vaccines, including Pfizer-BioNTech, Moderna, AstraZeneca, and Johnson Johnson, and has been administering them to eligible populations since December 2020.

One critical question in the fight against Covid-19 is the effectiveness of the vaccines in different age groups. It is well-known that age is a significant risk factor for severe illness, hospitalization, and death due to Covid-19, and older adults have been prioritized in vaccination campaigns globally.

However, there is limited information on how well the vaccines work among younger age groups, including children and Adults. Moreover, there have been concerns about the emergence of new Covid-19 variants, which could affect the effectiveness of the vaccines.

Understanding the effectiveness of Covid-19 vaccines among different age groups is essential for guiding public health policies and vaccine distribution strategies. Therefore, this study aims to investigate the effectiveness of Covid-19 vaccines in Saskatchewan among different age groups.The data collected from the Saskatchewan Health Authority were used to analyze the effectiveness of vaccines in preventing Covid-19 cases among different age groups.

This study has significant implications for public health policy and vaccine distribution in Saskatchewan and beyond. It provide valuable insights into the effectiveness of Covid-19 vaccines among different age groups and help guide vaccination strategies to achieve optimal protection against Covid-19.

1.2 Motivation

The Covid-19 pandemic has had a significant impact on the health and well-being of people worldwide, including in Saskatchewan. The development and distribution of Covid-19 vaccines have been a significant breakthrough in the fight against the pandemic. However, the effectiveness of the vaccines in different age groups is still an important research question, as age is a significant risk factor for severe illness and mortality due to Covid-19.

Saskatchewan has been administering several Covid-19 vaccines since December 2020, and there is a need to understand the effectiveness of these vaccines in different age groups. This information is helpful in guiding vaccine distribution and public health policies to achieve optimal protection against Covid-19. Moreover, with the emergence of new Covid-19 variants, understanding the effectiveness of vaccines in different age groups is crucial in adapting to changing circumstances.

Therefore, the motivation for this research is to investigate the effectiveness of Covid-19 vaccines in Saskatchewan among different age groups, including children (1), Young Adults (2), Middle Aged Adults (3), Late Middle Aged Adults (4), Early Seniors (5), Mid Senior (6),Late Seniors (7), and Elderly (8). This study aims to provide valuable insights into the effectiveness of vaccines in preventing Covid-19 cases among higher risking age groups. The findings of this study can then be used to inform public health policies and vaccine distribution strategies to combat the Covid-19 pandemic.

1.3 Objectives

The study aims to achieve the following specific objectives:

1. The impact of vaccines on rate of Covid-19 Cases among different age groups in Saskatchewan, and how do these outcomes differ from those

who have not been vaccinated?

2. How the effectiveness of Covid-19 vaccine vary by age group in Saskatchewan, and what implications does this have for vaccine distribution and administration?

3. The association between age and the effectiveness of Covid-19 vaccines in preventing infection.

4. Describe how the multinomial logistic regression model work for this specific setting.

5. To provide recommendations for vaccine distribution and public health policies based on the findings of the study.

6. Which age groups are most likely to be unvaccinated, partially vaccinated, or fully vaccinated?

7. How accurate is the multinomial logistic regression model in predicting the risk of Covid-19 infection based on age group and vaccination status.

1.4 Research Statement

This research aims to assess the effectiveness of Covid-19 vaccines by comparing the probability of Covid-19 cases between vaccinated (fully or partially vaccinated) and unvaccinated groups in different age categories. The analysis will take into account the interaction effect of the rate of cases, considering the intensity of occurrence either higher or lower within each group.

1.5 Scope

Multinomial logistic regression model for analyzing vaccine effectiveness among different age groups is an important statistical model because vaccines can have different levels of effectiveness in different age groups based on there status of vaccination. Multinomial logistic regression model help identify the factors that are associated with different levels of vaccine effectiveness in each age group and provide insights into the most effective vaccination strategies for each age group of the population. By analyzing the data using a multinomial logistic regression model, the results identify which vacination status cases are most strongly associated with vaccine effectiveness and thus this information is used to optimize vaccination strategies for different age groups in the populations.

1.6 Summary

The Covid-19 pandemic has been caused severe global health crisis, affecting millions of people worldwide. In response to this crisis, scientists and researchers have been working hard to develop and distribute vaccines to protect people from the virus.

While the vaccines have been shown to be effective in preventing Covid-19, it is important to understand how effective they are among different age groups. This is because different age groups may have varying levels of immune response to the vaccine, and may also have different levels of exposure to the virus.

To understand the effectiveness of the vaccines among different age groups, a statistical model called multinomial logistic regression modeling is used. This modeling technique allows to analyze the relationship between categorical variables, such as vaccine effectiveness and age group, while taking into account other relevant factors like rate of cases, type of cases , gender, comorbidities, and vaccine type.

By using multinomial logistic regression modeling, the probability of vaccine effectiveness among different age groups are identified, and also any factors that may influence vaccine effectiveness are determined. This information then be used to develop targeted vaccination strategies that are tailored to the needs of specific age groups, and will help to protect vulnerable populations from Covid-19.

In summary, understanding the effectiveness of Covid-19 vaccines among different age groups is critical in the ongoing efforts to combat the pandemic. Multinomial logistic regression modeling is a powerful tool that can help to analyze vaccine effectiveness in a comprehensive and meaningful way, and provide insights that are helpful for public health policies and interventions.

Chapter 2

Literature Review

2.1 Introduction

The Covid-19 pandemic has affected millions of people worldwide, leading to significant morbidity and mortality. Vaccines are a critical tool for controlling the spread of the virus and reducing its impact on public health. However, the effectiveness of Covid-19 vaccines varies among different age groups, with some groups showing higher vaccine effectiveness than others. In this literature review, we will examine the existing research on vaccine effectiveness among different age groups using multinomial logistic regression model.According to a literature review and meta-analysis by Abadie et al. (2021) Covid-19 vaccines have shown real-world effectiveness.

2.2 Covid-19 Vaccines Effectiveness

2.2.1 Public health Vaccine Effectiveness

Public Health Agency of Canada (2022) conducted analysis on the effectiveness of the vaccines on alpha and gamma variant in March 2021 and found that:

- A single dose of vaccine (mRNA or AstraZeneca/COVISHIELD) protected well against hospitalization, reducing the risk by more than 80%, including among age groups at highest risk of severe outcomes (60–69-year-olds and those 70 years and older).
- A single dose of a Covid-19 mRNA vaccine reduced the risk of getting Covid-19 by two-thirds in adults 70 years of age and older during the peak of the spring 2021 wave in Saskatchewan when Alpha (B.1.1.7) and Gamma (P.1) variants made up about 70% of circulating strains.
- Risk reduced by 80% in long-term care residents and health care workers, the first people to be vaccinated in Saskatchewan. There was also a reduction in hospitalizations and deaths among vaccinated long-term care residents.

Furthermore, the Public Health Agency of Canada (2022) also conducted an analysis on the effectiveness of the vaccines against the Delta variant in October 2021 and found that:

1. Two doses prevented about 95% of Covid-19 hospitalizations.

2. Two doses of the AstraZeneca vaccine were more than 70% effective against Covid-19 infection. People who got one AstraZeneca dose followed by one mRNA vaccine dose (also called “mix and match”) had protection that was as good as with two mRNA doses.

3. Vaccine protection was stronger when people received their second dose more than six weeks after their first dose.

In addition, an analysis of the Omicron variant on the effectiveness of the vaccines by the Public Health Agency of Canada revealed that :

1. Two doses provided good protection against severe outcomes.

2. Two doses prevented about 65-75% of hospitalizations (reducing the risk of Covid-19 hospitalization by about two-thirds to three-quarters compared to unvaccinated people).

3. Two doses were less effective against Omicron infection (less than 10-15% against any infection).

2.2.2 Monitoring Incidence of Covid-19 Cases By Vaccination Status

This study by Scobie et al. (2021) gives findings from the crude analysis of surveillance data in which the results are consistent with recent studies reporting decreased VE (Vaccination Effectiveness) against confirmed infection but not hospitalization or death, during a period of Delta variant predominance and potential waning of vaccine-induced population immunity in US. In 13 U.S. jurisdictions, rates of Covid-19 cases, hospitalizations, and deaths were substantially higher in persons not fully vaccinated compared with those in fully vaccinated persons, similar to findings in other reports. After the week of June 20, 2021, when the SARS-CoV-2 Delta variant became predominant, the percentage of fully vaccinated persons among cases increased more than expected for the given vaccination coverage and a constant VE (Vaccination Effectiveness) . The IRR (Relative Risk ratios) for cases among persons not fully vaccinated versus fully vaccinated decreased substantially; IRRs (Relative Risk Ratios) for hospitalizations and deaths changed less overall, but moderately among adults aged greater than and equal to 65 years.

2.2.3 Covid-19 Cases and Vaccination Rates

The study by Cerio et al. (2021) indicate that variants may impact vaccine effectiveness, current vaccination efforts are helping forestall some cases in Newyork. Widespread vaccination is still an important goal. Primary care providers, public officials, and public health scientists should continue to urgently promote and support vaccination efforts.

2.2.4 Efficacy of Covid-19 Vaccines

Another study Yelin et al. (2021) on Associations of the BNT162b2 Covid-19 vaccine effectiveness with patient age and comorbidities shows that Quantifying real-world vaccine effectiveness, including both biological and behavioral effects, this analysis provides initial measurement of vaccine effectiveness across demographic groups.

2.3 Covid-19 Vaccines effectiveness among age groups

Several studies have been conducted globally to analyze the effectiveness of Covid-19 vaccines among different age groups. One such study was conducted by Abu-Raddad et al. (2021), who analyzed the effectiveness of the Pfizer-BioNTech vaccine in Qatar among people aged 16 years and older. The study found that the vaccine was highly effective in preventing Covid-19 in all age groups, with an overall effectiveness of 89.5%.

Another study by Shah et al. (2021) Bego analyzed the effectiveness of the Moderna vaccine among different age groups in the United States. The study found that the vaccine was highly effective in preventing Covid-19 in all age groups, with an overall effectiveness of 94.1%. However, the study also found that the vaccine was less effective in people aged 65 years and older, with an effectiveness rate of 86.4%.

A study by Hall et al. (2021) analyzed the effectiveness of the Pfizer-BioNTech vaccine among different age groups in Israel. The study found that the vaccine was highly effective in preventing Covid-19 in all age groups, with an overall effectiveness of 94%. However, the study also found that the vaccine was less effective in people aged 85 years and older, with an effectiveness rate of 89%.

Using multinomial logistic regression modeling, a study by Pawelek et al. (2021) analyzed the effectiveness of Covid-19 vaccines among different age groups in the United States. The study found that the vaccine was highly effective in preventing Covid-19 in all age groups, but that vaccine effectiveness varied by age. The study found that vaccine effectiveness was highest in people aged 16-24 years, with an effectiveness rate of 96.2%, and lowest in people aged 75 years and older, with an effectiveness rate of 85.7%.

In conclusion, these studies suggest that Covid-19 vaccines are highly effective in preventing Covid-19 cases in all age groups, but that vaccine effectiveness may vary by age. Multinomial logistic regression modeling can be used to analyze the effectiveness of the vaccines among different age groups, and can provide insights that can inform public health policies and interventions.

2.4 Multinomial Logistic regression Model

2.4.1 Application of MLR Model in Various field

Predicting College Enrollment Decision Using Multinomial Logistic Regression

Yoo and Kim (2016) conducted study for Predicting college enrollment decision using multinomial logistic regression. This study aims to predict the college enrollment decision of high school students based on their demographic characteristics, academic achievements, and attitudes towards college. The authors use multinomial logistic regression to model the relationship between the predictors and the three possible outcomes of college enrollment decision (enroll in a four-year college, enroll in a two-year college, or not enroll in college).

Multinomial Logistic Regression for Predicting Protein Structural Classes

Chen et al. (2017) used Multinomial logistic regression model for predicting protein structural classes. This study proposes a multinomial logistic regression model to predict the structural class of proteins based on their amino acid sequences. The authors use a dataset of protein sequences with known structural classes to train the model, and evaluate its performance on a test dataset.

Explaining and Predicting Technology Adoption: A Multinomial Logistic Regression Model

Hillel (2017) conducted this study in which he develops a multinomial logistic regression model to explain and predict the adoption of technology by individuals. The model considers factors such as age, gender, education, income, and attitudes towards technology, and predicts the likelihood of an individual adopting technology at different levels (early adopter, average adopter, or late adopter).

A Multinomial Logistic Regression Model for Predicting Credit Ratings

Sarkar (2018) conducted this study to propose a multinomial logistic regression model to predict the credit ratings of companies based on their financial and non-financial characteristics. The model considers variables such as profitability, leverage, asset quality, industry sector, and geographic location, and predicts the probability of a company being assigned one of several credit ratings.

Multinomial Logistic Regression with Application to Customer Satisfaction Analysis

Bhattacharyya and Chatterjee (2019) used Multinomial logistic regression with application to customer satisfaction analysis. This study applies multinomial logistic regression to analyze customer satisfaction data from a survey.

The authors develop a model to predict the probability of each level of customer satisfaction (dissatisfied, somewhat satisfied, and very satisfied) based on customer demographics, product usage patterns, and other variables.

Multinomial Logistic Regression for Predicting Cancer Subtypes

Li and Wang (2019) used Multinomial logistic regression for predicting cancer subtypes. This study uses multinomial logistic regression to predict the subtypes of cancer based on gene expression data. The authors develop a model to classify cancer samples into one of several subtypes based on the expression levels of a set of genes, and evaluate its performance on a dataset of breast cancer samples.

MLR Analysis of Factors Associated with Underweight, Overweight, and Obesity Among Urban Women

Ahmed and Shafiullah (2019) used Multinomial logistic regression analysis of factors associated with underweight, overweight, and obesity among urban women in Bangladesh. This study uses multinomial logistic regression to analyze the factors associated with underweight, overweight, and obesity among urban women in Bangladesh. The authors examine the relationships between various socioeconomic and demographic variables and the three weight categories, and identify the factors that are most strongly associated with each category.

A Multinomial Logistic Regression Approach to Predicting International Tourism Demand

Ertugrul and Karatas (2019) conducted this study to propose a multinomial logistic regression model to predict the international tourism demand for Turkey. The authors use data on the number of tourists from different countries and their demographic and travel characteristics to develop a model that predicts the probability of each country being a source of tourism demand for Turkey.

2.4.2 Application of MLR Model in Vaccine Efficacy

Acceptance of Vaccination

Zhang et al. (2014) conducted a study among medical students in China to analyze factors affecting acceptance of pandemic A/H1N1 influenza vaccination. They used multinomial logistic regression analysis and found that age, gender, and educational level were significant factors associated with vaccine acceptance.

Pneumococcal Vaccine Effectiveness

Another study by Yao et al. (2020) investigated the impact of the 23-valent pneumococcal polysaccharide vaccine using multinomial logistic regression model. The study found that the vaccine was highly effective in preventing pneumococcal infection among older adults aged 65 years and older.

Influenza Vaccine Effectiveness

Several studies have investigated vaccine effectiveness against influenza among older adults using multinomial logistic regression model.The study by Nguyen et al. (2021) conducted in which the results shows that vaccine effectiveness against influenza in older adults gives a nested case-control research in Quebec, Canada, and found that influenza vaccine effectiveness was highest among older adults aged 65 years and older.The study Sharma et al. (2019) conducted in a hospital-based case-control study in Italy found that inactivated influenza vaccines were more effective in preventing influenza infection among the elderly population.

Covid-19 Vaccination

Most recently, study Kodera et al. (2022) conducted in Japan to analyze the factors associated with willingness to receive Covid-19 vaccination among the general public. They used multinomial logistic regression analysis and found that age and trust in government were significant factors associated with vaccine acceptance among population.

2.5 Conclusion

The studies reviewed above demonstrate the usefulness of multinomial logistic regression analysis in analyzing vaccine effectiveness among different age groups. These studies provide valuable insights into the factors that influence vaccine effectiveness, which is used in this study to help guide public health policies and vaccination strategies makers. This review suggests that age is an important factor that affects vaccine effectiveness and acceptance among different populations, including the adult as well as elderly population. Therefore, it is crucial to understand the age-related factors associated with Covid-19 vaccine effectiveness to ensure the success of vaccination programs.

Chapter 3

Methodology

3.1 Introduction

The purpose of this study is to examine the effectiveness of Covid-19 vaccines among different age groups in Saskatchewan using a multinomial logistic regression model. The increasing prevalence of Covid-19 has made it crucial to understand the impact of vaccines on reducing the spread of the disease. This methodology chapter provide detail on the research questions, experimental design, data collection, and data analysis techniques used in this study.

3.2 Research Question

This study is conducted to answer following questions:

1. What is the relationship between age group, vaccination status, and

the incidence of Covid-19 cases, and how do these variables interact in predicting the risk of infection?

2. What are the predicted probabibility of risk of Covid-19 cases with different vaccination status among different age groups in Saskatchewan, and how do these differ from those who have not been vaccinated?

3. How does the effectiveness of Covid-19 vaccine vary by age group in Saskatchewan, and what implications does this have for vaccine distribution and administration?

4. To what extent does the age group of individuals in Saskatchewan affect their likelihood of getting vaccinated against Covid-19, and how does this vary by vaccination status?

3.3 Experimental Design

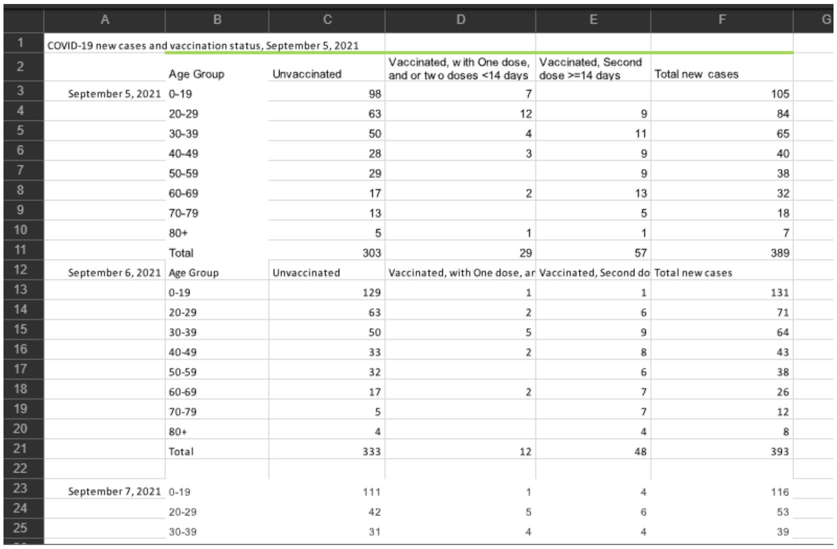
The experimental design for this study is a retrospective, observational cohort study using publicaly available data obtain from the official website of Government of Saskatchewan Vaccination coverage. The study population consist of individuals of different age groups who have Covid-19 along with three different vaccination status i.e unvaccinated cases, partially vaccinated cases and fully vaccinated cases in Saskatchewan collected between September 2021 and December 2021. The primary outcome measure are the odd ratios obtained by using multinomial logistic regression model and then by using these odd ratios effectiveness of Covid-19 vaccines among different age groups in Saskatchewan was fond which indicate reduction in symptomatic cases among vaccinated individuals. Also this study used the research paper 'Goodness of Fit of Product Multinomial Regression Models to Sparse Data' by Deng and Paul (2003) as a reference to derived the mathematical proof of multinomial logistic regression model and also development of this model with interaction effect to find the effectiveness of Covid-19 Vaccines. Specifically, this study incorporated the research methods and techniques for assessing the goodness of fit of this study's multinomial logistic regression model which helped to ensure the validity and reliability of results.

3.4 Data Collection

Data for this research is collected on the daily basis from the official website of Government of Saskatchewan that shows the number of case with their specific vaccination status. The data is collected for different age groups during a period of four Months.

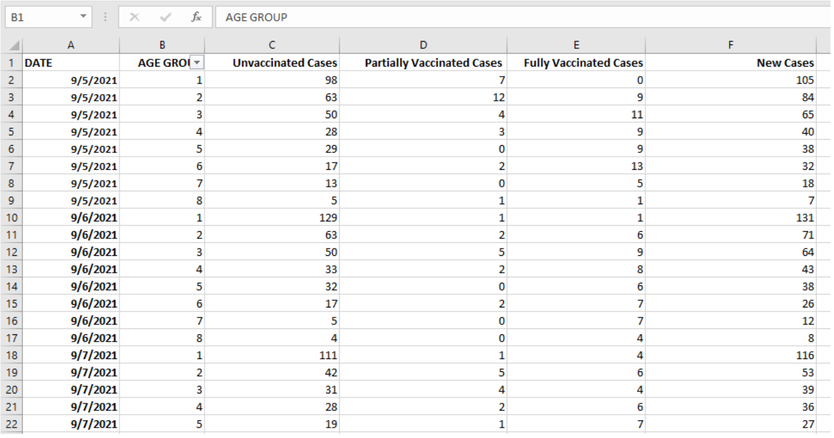
3.4.1 Number of cases reported

Covid-19 cases were reported per 100,000 population in saskatchewan. According to a report sas, the number of Covid-19 cases in Saskatchewan On September 24, 2021 was : Of the 528 new Covid-19 cases , 426 (81%) unvacci-



	A	B	C	D	E	F	G
1	COVID-19 new cases and vaccination status, September 5, 2021						
2		Age Group	Unvaccinated	Vaccinated, with One dose, and or two doses <14 days	Vaccinated, Second dose >=14 days	Total new cases	
3	September 5, 2021	0-19	98	7		105	
4		20-29	63	12	9	84	
5		30-39	50	4	11	65	
6		40-49	28	3	9	40	
7		50-59	29		9	38	
8		60-69	17	2	13	32	
9		70-79	13		5	18	
10		80+	5	1	1	7	
11		Total	303	29	57	389	
12	September 6, 2021	Age Group	Unvaccinated	Vaccinated, with One dose, ar	Vaccinated, Second do	Total new cases	
13		0-19	129	1	1	131	
14		20-29	63	2	6	71	
15		30-39	50	5	9	64	
16		40-49	33	2	8	43	
17		50-59	32		6	38	
18		60-69	17	2	7	26	
19		70-79	5		7	12	
20		80+	4		4	8	
21		Total	333	12	48	393	
22							
23	September 7, 2021	0-19	111	1	4	116	
24		20-29	42	5	6	53	
25		30-39	31	4	4	39	

Figure 3.1: Original data file



B1 | AGE GROUP

	A	B	C	D	E	F
1	DATE	AGE GROU	Unvaccinated Cases	Partially Vaccinated Cases	Fully Vaccinated Cases	New Cases
2	9/5/2021	1	98	7	0	105
3	9/5/2021	2	63	12	9	84
4	9/5/2021	3	50	4	11	65
5	9/5/2021	4	28	3	9	40
6	9/5/2021	5	29	0	9	38
7	9/5/2021	6	17	2	13	32
8	9/5/2021	7	13	0	5	18
9	9/5/2021	8	5	1	1	7
10	9/6/2021	1	129	1	1	131
11	9/6/2021	2	63	2	6	71
12	9/6/2021	3	50	5	9	64
13	9/6/2021	4	33	2	8	43
14	9/6/2021	5	32	0	6	38
15	9/6/2021	6	17	2	7	26
16	9/6/2021	7	5	0	7	12
17	9/6/2021	8	4	0	4	8
18	9/7/2021	1	111	1	4	116
19	9/7/2021	2	42	5	6	53
20	9/7/2021	3	31	4	4	39
21	9/7/2021	4	28	2	6	36
22	9/7/2021	5	19	1	7	27

Figure 3.2: Modified data file

nated, 27 (5%) were partially vaccinated and 75 (14%) were fully vaccinated. On December 3, 2021 was : Of the 78 new Covid-19 cases , 52 (18 per 100,000) (67%) unvaccinated, 2 (2.6 per 100,000) (2%) were partially vaccinated and 24 (2.9 per 100,000) (31%) were fully vaccinated.

3.4.2 Data Collection Method and Technique

The data for this study was collected through daily monitoring of the official website of the Government of Saskatchewan. The website provides the number of Covid-19 cases with their specific vaccination status for different age groups.

3.4.3 Vaccination Coverage Plan

According to a report on the Covid-19 vaccination rollout in Saskatchewan Rowein et al. (2022), the province has made progress in vaccinating its population.

The vaccination coverage plan is based on the eligibility of age, with different priority groups defined within each age range. This plan aims to ensure that those who are most at risk of severe illness or exposure to Covid-19 are vaccinated first.The summary of Vaccine Eligibility plan is given in table 3.1.

Table 3.1: Summary of Vaccine Eligibility by Priority Group

Priority Group	**Eligible phase of vaccine roll-out**	**Date eligible**	**Other notes**
Children, <12 years	Phase 2	Nov 24, 2021	
Youth, 12-18 years	Phase 2	May 20, 2021	Pfizer-BioNTech (12-16), Moderna (16-18)
Young adults	Phase 2	May 2021	Moderna (ages 18-29)
Adults	Phase 2	Apr-May 2021	Ages 30-59; Also eligible for AstraZeneca Vaxzevria® (only group eligible for the Astra Zeneca vaccine between Apr 28 - May 6, 2021; discontinued for first doses as of May 6)
Older adults	Phase 2	Mar 2021	Mar 18: 67+; Mar 24: 65+; Mar 25: 62+; Mar 31: 60+
Seniors	Phase 1	Dec 22, 2020	Residents 70+ in all communities

3.5 Data Analysis

Data are analyzed using a multinomial logistic regression model to determine the association between age and vaccine effectiveness. The model adjust for potential confounders, including demographic characteristic i.e age groups. The odds ratios from the model is used to estimate the relative risk/ odd ratios of vaccine effectiveness for different age groups.

3.5.1 Key Variables

The key variables that have impact on the outcomes of interest are given in table 3.2. The table provides descriptions of variables including dependent variables such as vaccination status and independent variables such as age groups and Covid-19 cases. Each variable is associated with levels or range of values which are useful for analyzing the impact of these variables on the outcomes of interest. As, data was collected for eight different age groups instead of sorting them based on who got vaccinated according to vaccination coverage plan, this study used an approximation method to understand the data from these eight groups better, as they gave more detailed information about each group.Whereas, an approximation method refers to using the available data from the eight different age groups to make an estimate about the vaccination coverage for each group, even though the data was not specifically sorted based on vaccination status. This information is used to get a general idea of the vaccination coverage for each age group.

Table 3.2: Descriptions of variables

Variable Type	**Variable Name**	**Variable Description**	**Levels or Range of Values**
Dependent Variables	Vaccination Status	Number of cases reported with different statuses	1 = unvaccinated case, 2 = partially vaccinated case, 3 = Fully vaccinated case
Independent Variables	Age groups	Cases reported according to age group	1 = 0-19, 2 = 20-29, 3= 30-39, 4= 40-49, 5= 50-59, 6= 60-69, 7= 70-79, 8= 80+
	Covid-19 Cases	Case count for each status	0= no case , 1= lower number of cases, 2= higher number of cases

3.5.2 Data Pre-processing

The data preprocessing has been performed on the dataset to prepare it for analysis. This includes:

- Data Cleanup : Multiple imputation technique is used to removed any missing data in the dataset. This method involves creating multiple imputed datasets by randomly replacing missing values with plausible values based on the observed data. The regression model is then run on each imputed dataset, and the results are combined to generate final estimates and standard errors. This method is used because it produce more accurate results.

- Outlier detection and removal: Outliers are checked within the dataset that may affect the accuracy of the model. Boxplots visualizations are used to identify outliers and then remove them from the dataset.

- Data encoding: Categorical variables are encoded. The Vaccination Status, Rate of Cases, and Age variables are converted to factors.

3.5.3 Limitations

Potential limitations of this study include missing data, selection bias, small sample size and confounding variables. Missing data are handled by applying multiple imputation technique. Selection bias are minimized by using random sampling techniques. Small sample size represent the population and gives the approximation results for entire population. Confounding variables are controlled by adjusting for potential confounding factors in the multinomial logistic regression model.

3.6 Summary

Data for this research was collected through daily monitoring of the official website of the Government of Saskatchewan. This website provided the number of Covid-19 cases with specific vaccination status for different age groups over a four-month period i.e September 2021 to December 2021. The data collection process involved manually recording the number of cases for each status and age group on a daily basis. To relate the vaccination plan to vaccination status, the data are analyzed on the eligibility criteria for the vaccine rollout in Saskatchewan. This eligibility was based on age, with different priority groups defined within each age range. By comparing the vaccination status of individuals with their eligibility for the vaccine, we are able to analyze the effectiveness of the vaccination plan used for providing protection against Covid-19. We also identified key variables that have an impact on the outcomes of interest, including vaccination status and Covid-19 cases. These variables were associated with levels or ranges of values, such as the different vaccination status categories and age groups. This helps in analyzing the impact of these variables on the outcomes of interest and provide insights into the effectiveness of the vaccination plan in Saskatchewan. In conclusion, the methodology used in this study includes a retrospective, observational cohort study using a multinomial logistic regression model. The study has been designed to minimize potential limitations and to ensure valid results.

Chapter 4

Theory

4.1 Multinomial Random Variable

The multivariate discrete random variable $Y = (Y_1, \cdots, Y_P)$ is said to follow the multinomial distribution with multinomial denominator m and multivariate probability $\pi = (\pi_1, \ldots, \pi_p)$ if

$$P(Y_1 = y_1, \ldots, Y_p = y_p) = \frac{m!}{y_1! \cdots y_p!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_p^{y_p} \tag{4.1}$$

where $y_1 + y_2 + \cdots + y_p = m$ and $\pi_1 + \pi_2 + \cdots + \pi_p = 1$. Further, the mean vector of Y and covariance matrix is

$$\mu = (\mu_1, \ldots, \mu_p) = (m\pi_1, \ldots, m\pi_p) \tag{4.2}$$

4.1.1 Properties of multinomial random variable

1. Definition: A multinomial random variable is defined as the number of times that each category occurs in a set of n independent trials, where each trial results in the object being placed into one of k mutually exclusive and exhaustive categories.

2. Probability mass function: The probability mass function of a multinomial random variable gives the probability that the variable takes on a particular set of values. If X is a multinomial random variable with parameters m and $(p_1, p_2, \cdots, p_k)$, then the probability mass function is given by:

$$P(X = (x_1, x_2, \ldots, x_k)) = \frac{m!}{(x_1!\, x_2! \ldots x_k!\,)} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \tag{4.3}$$

   where $x_1 + x_2 + \ldots + x_k = m$ and p is the probability of category i in each trial.

3. Mean, variance and covariance: The mean, variance and covariance of a multinomial random variable are given by:

$$E(X) = \mu = (\mu_1, \cdots, \mu_p) = (mp_1, mp_2, \ldots, mp_k) \tag{4.4}$$

$$V(X) = (mp_1(1 - p_1), mp_2(1 - p_2), \ldots, mp_k(1 - p_k)) \tag{4.5}$$

$$Cov(X) = E[(X-\mu)(X-\mu)^T] \tag{4.6}$$

where $\mu = (\mu_1, \mu_2, ..., \mu_k)$ is the vector of expected values or means of X, with $\mu_i = E[X_i] = mp_i$, where m is the total number of observations.

4. Independence: The counts in each category of a multinomial random variable are independent if the trials are independent.

5. Multinomial distribution: The joint distribution of the counts in each category of a multinomial random variable is a multinomial distribution, and the marginal distribution of the count in any individual category is a binomial distribution.

6. Applications: Multinomial random variables are commonly used in fields such as statistics, biology, and marketing research to model experiments in which objects are placed into multiple categories.

4.2 Definition of Multinomial logistic regression model

Multinomial logistic regression is a statistical model used to analyze and predict the relationship between a categorical dependent variable with more than two categories (i.e., a nominal or ordinal variable with three or more categories) and one or more independent variables (predictor variables) that can be either categorical or continuous. It is an extension of binary logistic regression, which is used when the dependent variable has only two categories. Multinomial logistic regression provides estimates of the probability of each category of the dependent variable occurring given the values of the independent variables. Multinomial logistic regression distribution is a probability distribution used to model a categorical dependent variable with more than two categories (i.e., a nominal or ordinal variable with three or more categories) in a regression analysis. It assumes that the dependent variable follows a multinomial distribution, which describes the probability of observing each category of the dependent variable given a set of predictor variables.

The model estimates the probability of each category of the dependent variable based on a linear combination of the predictor variables, which can be either categorical or continuous. The coefficients estimated for each predictor variable represent the change in the log-odds of being in a particular category of the dependent variable for a unit change in that predictor variable while holding all other predictors constant.

Multinomial logistic regression is a commonly used method for analyzing data in a variety of fields such as social sciences, public health, and economics. It is particularly useful when the dependent variable has more than two categories and there is interest in understanding the relationship between the dependent variable and one or more predictor variables. Multinomial logistic regression uses the multinomial logistic distribution, also known as the softmax distribution, for modeling the probabilities of the dependent variable categories. The multinomial logistic distribution is a generalization of the logistic distribution that can handle multiple categories. It is a type of probability distribution that assigns probabilities to a set of mutually exclusive events, such as the possible outcomes of a categorical variable with k categories. The probabilities for each category are obtained by applying the softmax function to the linear combination of the predictor variables, which transforms the logit into a set of k probabilities that sum to 1.

4.3 Components of Multinomial logistic regression model

1. Categorical dependent variable: The dependent variable Y in a multinomial logistic regression model is categorical and can take on k different categories: $Y = (y_1, y_2, \ldots, y_k)$, where $k \geq 2$.

2. Linear relationship between predictors and logits: The log-odds of the dependent variable categories are modeled as a linear combination of the predictor variables, denoted by $(X_1, X_2, \ldots, X_p)$:
   $\text{logit}(p(y_i = 1)) = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \ldots + \beta_{1p}X_p = X^T\beta_1$
   $\text{logit}(p(yi = 2)) = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \ldots + \beta_{2p}X_p = X^T\beta_2$
   $\ldots$
   $\text{logit}(p(yi = k)) = \beta_{k0} + \beta_{k1}X_1 + \beta_{k2}X_2 + \ldots + \beta_{kp}X_p = X^T\beta_k$

3. Maximum likelihood estimation: The model parameters (β coefficients) are estimated using the maximum likelihood estimation method, which

finds the values of the parameters that maximize the likelihood of the observed data.

4. Goodness-of-fit tests: Various goodness-of-fit tests can be used to assess the model fit, including the likelihood ratio test, the deviance goodness-of-fit test, and the Pearson chi-square test.

5. Multicollinearity: Multinomial logistic regression assumes that there is no multicollinearity among the predictor variables. That is, the predictor variables should not be highly correlated with each other.

6. Interpretation of coefficients: The coefficients in the model represent the change in the log-odds of the dependent variable categories for a unit change in the corresponding predictor variable, while holding all other variables constant. The coefficients can be exponentiated to obtain odds ratios, which represent the change in odds of the dependent variable categories for a unit change in the corresponding predictor variable such as

$$OR(y_i = 1) = \exp(\beta_{1j}), \ldots, OR(y_i = k) = \exp(\beta_{kj}) \tag{4.7}$$

where j denotes the j^{th} predictor variable.

4.3.1 Properties of MLR Model

The multinomial logistic regression model is a powerful statistical tool used to predict the probability of a categorical outcome variable with three or more categories. The key properties of the multinomial logistic regression model include:

1. Multinomial outcome variable: The outcome variable in the multinomial logistic regression model is multinomial, meaning it has three or more categories.

2. Predictive power: The multinomial logistic regression model is a powerful predictor of the probability of the outcome variable, based on the values of the predictor variables.

3. Interpretability: The coefficients of the multinomial logistic regression model can be interpreted in a similar way to the coefficients in a binary logistic regression model, allowing for the identification of the direction and strength of the relationship between the predictor variables and the outcome variable.

4. Goodness of fit: The multinomial logistic regression model can be used to assess the goodness of fit of the model to the data, which allows for the evaluation of the model's overall performance in predicting the outcome variable.

5. Model comparison: The multinomial logistic regression model can be

used to compare the relative performance of different models with different predictor variables, allowing for the identification of the most important predictor variables in predicting the outcome variable.

6. Flexibility: The multinomial logistic regression model can accommodate a wide range of predictor variables, including continuous, categorical, and binary variables.

7. Model assumptions: The multinomial logistic regression model is based on certain assumptions, including linearity, independence of observations, no multicollinearity, and homogeneity of variance.

4.3.2 Assumptions of MLR Model

The multinomial logistic regression model is a type of regression analysis used to predict the probability of a categorical outcome variable with three or more categories. The assumptions of the multinomial logistic regression model include:

1. Independence of observations: The observations are assumed to be independent of each other.

2. Linearity: The relationship between the predictor variables and the log-odds of the outcome variable should be linear.

3. No multicollinearity: The predictor variables should not be highly correlated with each other.

4. Large sample size: The sample size should be large enough to ensure stable estimates of the regression coefficients and standard errors.

5. Absence of outliers: The data should not contain extreme values that may unduly influence the model results.

6. Homogeneity of variance: The variance of the log-odds should be constant across the levels of the outcome variable.

7. No perfect separation: The data should not exhibit perfect separation, which occurs when there is a combination of predictor variables that perfectly distinguishes between the different levels of the outcome variable.

8. Independence of irrelevant alternatives: The model should not be affected by the addition or removal of an irrelevant alternative category.

4.4 Logistic regression Model

The logistic model is used to model the probability of an event occurring. For this, we have k events, and we want to model the probability of each event i relative to the k^{th} event. The probability of event i relative to the k^{th} event is represented by $\frac{\pi_i}{\pi_k}$. The logistic model assumes that the log odds ratio of each event relative to the k^{th} event is a linear function of covariates

X, with parameters β_i. This can be written as:

$$\log\left(\frac{\pi_i}{\pi_k}\right) = X\beta_i \quad ; i = (1, 2, \ldots, k-1) \tag{4.8}$$

4.4.1 Logistic Regression Model with Covariates

In logistic regression, covariates refer to the independent variables or predictor variables that are used to explain the relationship between a binary dependent variable and one or more independent variables. The logistic regression model estimates the effect of the covariates on the probability of the binary outcome. The formula for the logistic regression model with one covariate is:
$\text{logit}(p) = \beta_0 + \beta_1 X$ where:

- ***logit*** is the natural logarithm of the odds of the binary outcome.
- ***p*** is the probability of the binary outcome.
- $\boldsymbol{\beta_0}$ is the intercept, which represents the log odds of the binary outcome when all covariates are equal to zero.
- $\boldsymbol{\beta_1}$ is the coefficient of the covariate X, which represents the change in log odds of the binary outcome for a one-unit increase in X.

The formula for the logistic regression model with multiple covariates is:
$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_p$ where $(X_1, X_2, \ldots, X_p)$ are the p covariates and $(\beta_1, \beta_2, \ldots, \beta_k)$ are the coefficients of the k covariates. Also,

β_0 is the intercept. The probabilities π_i must add up to 1. Additionally, we can represent π_k as 1 minus the sum of the other probabilities. These constraints can be written as:

$$\sum_{i=1}^{k} \pi_i = 1 \tag{4.9}$$

$$\pi_k = 1 - (\pi_1 + \pi_2 + \ldots + \pi_{k-1}) \tag{4.10}$$

Finally, the covariates X can be represented as a p-dimensional vector of covariates, with X_i representing the i^{th} covariate. Similarly, the parameters β_i can be represented as a p-dimensional vector of parameters for each event i. This can be written as:

$$X = (X_1, X_2, \ldots, X_p) \tag{4.11}$$

$$Y = (Y_1, Y_2, \ldots, Y_k) \tag{4.12}$$

$$\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip}) \tag{4.13}$$

$$X\beta_i = (X_1\beta_{i1}, X_2\beta_{i2}, \ldots, X_p\beta_{ip}) \tag{4.14}$$

Thus,

$$\log(\frac{\pi_i}{\pi_k}) = X\beta_i \quad where \quad i = (1, 2, \ldots, k-1) \tag{4.15}$$

In logistic regression, the coefficients represent the change in log odds of the binary outcome for a one-unit increase in the corresponding covariate, holding all other covariates constant. The exponentiated coefficients, called odds ratios, represent the multiplicative effect of the corresponding covariate on the odds of the binary outcome, relative to the reference category. In summary, covariates in logistic regression refer to the independent variables or predictor variables that explain the relationship between a binary dependent variable and one or more independent variables. The logistic regression model estimates the effect of the covariates on the probability of the binary outcome using coefficients, which represent the change in log odds of the binary outcome for a one-unit increase in the corresponding covariate, holding all other covariates constant.

4.5 Model Design

The model assumes that there are k categories and that each observation belongs to one and only one category. The response variable is represented by Y, and each observation i has a vector of categorical responses Y_i, where the elements of the vector represent the counts for each category.

4.5.1 Multinomial Distribution

The model consisits of multinomial random variables which mean that the vector of responses Y_i follows a multinomial distribution with parameters m_i and π_i. The parameter m_i represents the total number of trials or observations for observation i, and π_i is a vector of probabilities representing the probability of each category for observation i. The probabilities π_i are related to a linear combination of predictor variables Z and a vector of parameters β . Specifically, π_{ik} is the probability of observation i belonging to category k, and it is related to the linear combination $Z_i^T\beta$ through the link function. The link function is used to ensure that the probabilities π_i sum to 1 across all categories. The linear combination $Z_i^T\beta$ is also related to the expected value of Y_i through the use of the link function $h_i\eta_i$. The link function is used to model the relationship between the linear predictor $Z_i^T\beta_i$ and the expected value of Y_i, which is denoted as μ_i. The link function is typically chosen to ensure that μ_i is non-negative and has a specified range of values. For the multinomial model, the link function is usually chosen to be the logarithm function, and it is defined as follows:

$$\mu_i = m_i\pi_i = (m_i\pi_{i1}, m_i\pi_{i2}, \dots, m_i\pi_{ip}) \tag{4.16}$$

$$h_i\eta_i = (\eta_1(\mu_{i1}), \eta_2(\mu_{i2}), \dots, \eta_p(\mu_{ip})) = Z_i^T\beta \tag{4.17}$$

$Let \quad y_i \quad where \quad i = (1, 2, \dots, n) \; Also, \quad Y = (Y_1, Y_2, \dots, Y_k) \; Then, \quad Y_i = (y_{i1}, y_{i2}, \dots, y_{ip})$ Now,

$$Y_i \sim multinomial \quad (m_i, \pi_i) \tag{4.18}$$

where, $\pi_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{ip})$ and $\mu_i = m_i\pi_i = (m_1\pi_{i1}, m_2\pi_{i2}, \dots, m_i\pi_{ip})$

$$\mu_i = h_i(\eta_i) = (h_{i1}\eta_{i1}, h_{i2}\eta_{i2}, \dots, h_{ip}\eta_{ip}) \tag{4.19}$$

$$h_i(\eta_i) = (h_{i1}(Z_i^T\beta_1,\ldots,Z_i^T\beta_p), h_{i2}(Z_i^T\beta_1,\ldots,Z_i^T\beta_p),\ldots,h_{ip}(Z_i^T\beta_1\ldots,Z_i^T\beta_p)) \tag{4.20}$$

$$\pi_{i1} = \frac{e^{Z_i^T\beta_1}}{1+\Sigma e^{Z_i^T\beta_1}} \tag{4.21}$$

$$\pi_{ik} = \frac{e^{Z_i^T\beta_k}}{1+\Sigma_{h=1}^{p} e^{Z_i^T\beta_h}} \tag{4.22}$$

where $h = (1,2,\ldots,p-1)$, $\pi_{ip} = \frac{1}{1+\sum_{h=1}^{p} e^{Z_i\beta_h}}$ and $\Sigma_{h=1}^{p}\pi_{ip} = 1$ Also, β_s is the parameter of s^{th} component in $(Y_1, Y_2, \ldots, Y_s, \ldots, Y_k)$. For $g_r(\beta)$ Let

$$Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iq}) \tag{4.23}$$

$$\beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jq}) \tag{4.24}$$

$$Z_1^T\beta_j = \beta_{j1}Z_{i1} + \beta_{j2}Z_{i2} + \ldots + \beta_{jq}Z_{iq} = \Sigma_{h=1}^{q}\beta_{jh}Z_{ih} \tag{4.25}$$

$q^{'} = pq$ from the Deng and Paul (2003) is used which is the likelihood estimating equation and p-dimensional vector of equations :

$$g_r(\beta) = \sum_{i=1}^{n}(Y_i - \mu_i)^T \sum_{i}^{-1} \frac{\partial\mu_i}{\partial\beta_r} = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{Y_{ij}-\mu_{ij}}{\mu_{ij}} \frac{\partial\mu_{ij}}{\partial\beta_r} \tag{4.26}$$

$$g(\beta) = \begin{bmatrix} g_1(\beta) \\ g_2(\beta) \\ \vdots \\ g_p(\beta) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}(Y_{i1}-\mu_{i1})^T\Sigma_i^{-1}\frac{\partial\mu_{i1}}{\partial\beta_{1r}} \\ \sum_{i=1}^{n}(Y_{i2}-\mu_{i2})^T\Sigma_i^{-1}\frac{\partial\mu_{i2}}{\partial\beta_{2r}} \\ \vdots \\ \sum_{i=1}^{n}(Y_{ip}-\mu_{ip})^T\Sigma_i^{-1}\frac{\partial\mu_{ip}}{\partial\beta_{pr}} \end{bmatrix} \tag{4.27}$$

$$(\hat{\beta} - \beta_0)^T \sum^{-1} (\hat{\beta} - \beta_0) \sim \chi^2(pq) \tag{4.28}$$

Then the expression for covariance matrix pxq can be written as:

$$\sum = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1\times pq} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2\times pq} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{pq\times 1} & \sigma_{pq\times 2} & \cdots & \sigma_{pq}^2 \end{pmatrix} \tag{4.29}$$

where σ_i^2 is the variance of β_i, σ_{ij} is the covariance between β_i and β_j, and $p \times q$ is the size of the matrix. Now by using this we have $\beta_r = \beta_r^0$

$$(\hat{\beta}_r - \beta_r^0)^T \sum_r^{-1} (\hat{\beta}_r - \beta_r^0) \sim \chi^2(pq) \tag{4.30}$$

$$-\frac{\partial q_r(\beta)}{\partial \beta_r \partial \beta_r^T} = -\sum\sum \frac{\partial (Y_{ij} - \mu_{i1} \partial \mu_j)}{\partial \beta_r^T (\mu_{ij} \partial \beta_r)} \tag{4.31}$$

$I = \sum_r = \sigma_{rst} \quad where \quad s = (1, 2, \ldots, p)$ Also, $\sigma_{rst} = \sum_i \sum_j \frac{\partial (Y_{ij} - \mu_{ij})}{\partial \beta_{rt} (\mu_{ij} \partial \beta_{rs})}$

In case of dependent variable for this specific case we have $p = 3$ and $Cov(\hat{P}_r) = I^{-1}$

$$(\hat{\beta}_r - \beta_r^0) \quad [Cov(\beta_0)]^{-1} (\hat{\beta}_r - \beta_r^0) \tag{4.32}$$

$$(\hat{\beta}_r - \beta_r^0)I(\hat{\beta}_r - \beta_r^0) \sim \chi^2(pq) \tag{4.33}$$

4.5.2 Link Function

The link function for multinomial logistic regression model with three categories of dependent variables are :

$$\pi_{ij} = \frac{\exp(\beta_{j1} + \log(\alpha_i)\beta_{j2})}{(1 + \sum_{k=1}^{p-1} \exp(\beta_{k1} + \log(\alpha_i)\beta_{k2})} \tag{4.34}$$

where $\alpha_i = age(i)$ is the age groups and β is the vector of parameters. The expected value of Y_i is given by:

$$\mu_i = m_i\pi_i = (m_i\pi_{i1}, m_i\pi_{i2}, \ldots, m_i\pi_{ip}) \tag{4.35}$$

where π_i is the vector of probabilities for observation i.

4.5.3 Estimation Method

The estimation approach for the multinomial logistic regression model is maximum likelihood estimation. Maximum likelihood estimation is a statistical method used to estimate the parameters of a statistical model based on observed data. The goal of maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood of observing the data.

In the case of multinomial logistic regression, the likelihood function is the product of the probabilities of each observation belonging to its respective category. The parameters of the model are the coefficients of the predictor variables that are used to predict the probabilities of the categories. The maximum likelihood estimator is the set of coefficient values that maximize the likelihood function.

The estimation procedure involves calculating the log-likelihood function, which is the natural logarithm of the likelihood function, and then finding the maximum value of the log-likelihood function. The maximum likelihood estimator can be obtained by solving the first-order conditions of the log-likelihood function with respect to the coefficients. This is typically done using numerical optimization algorithms, such as the Newton-Raphson method or gradient descent.

The maximum likelihood estimator provides estimates of the coefficients that best fit the data and can be used to make predictions about the probabilities of the categories for new observations. The estimator also provides measures of the uncertainty in the parameter estimates, such as standard errors and confidence intervals. The parameters β are estimated using maximum likelihood estimation, which involves finding the values of β that maximize the likelihood function. The likelihood function is defined as the joint probability density function of the observed data given the parameters. In the case of the multinomial model, the likelihood function is given by:

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{K} P(Y_{ij} = j)^{I(Y_{ij}=j)} \tag{4.36}$$

The log-likelihood function is then:

$$\ell(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{K} I(Y_{ij} = j) \ln(P(Y_{ij} = j)) \tag{4.37}$$

Here, $\ell(\beta)$ is the log-likelihood function, $L(\beta)$ is the likelihood function, n is the number of observations, p is the number of predictors, Y_{ij} is the i^{th} observation of the j^{th} predictor variable, and π_{ij} is the predicted probability of the i^{th} observation belonging to the j^{th} class.

The score function for the multinomial model is the vector of partial derivatives of the log-likelihood function with respect to each parameter β_j:

$$\nabla \ell(\beta) = (g_1(\beta), g_2(\beta), \ldots, g_p(\beta)) \tag{4.38}$$

where $g_j(\beta)$ is the partial derivative of the log-likelihood function with respect to β_j. The score function is used in optimization algorithms to find the values of β that maximize the likelihood function.

4.5.4 Algorithm for Multinomial logistic regression model

After coding the nominal scale outcomes, a comparison of the interactions between the age groups, cases, and Vaccination Status are customized for the model to understand the relationship between the variables in the model. The baseline category was fully vaccinated cases and with the other categories being based against this baseline it corresponds to this model given below. A mathematical proof of the maximum likelihood estimation approach used by the nnet library to estimate the coefficients in a multinomial logistic regression model:

Suppose we have a dataset consisting of n observations, denoted as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is a p-dimensional vector of independent variables and y_i is a categorical dependent variable with k categories. The multinomial logistic regression model specifies the probability of observing the dependent variable as y_i, given the independent variables x_i, as:

$$P(y_i = j|x_i, \beta) = \frac{\exp(\beta_{j0} + \beta_{j1}x_{1i} + \beta_{j2}x_{2i} + ... + \beta_{jp}x_{pi})}{\sum_{l=1}^{k} \exp(\beta_{l0} + \beta_{l1}x_{1i} + \beta_{l2}x_{2i} + ... + \beta_{lp}x_{pi})} \tag{4.39}$$

where β_{j0} is the intercept term for the j^{th} category, β_{jp} is the coefficient for the p^{th} independent variable in the j^{th} category, and β is a vector of coefficients to be estimated.

The likelihood function of the multinomial logistic regression model is given by:

$$L(\beta|y, x) = \prod_{i=1}^{n} \prod_{j=1}^{k} [P(y_i = j|x_i, \beta)]^{I(y_i=j)} \tag{4.40}$$

where $I(y_i = j)$ is an indicator function that takes the value of 1 if $y_i = j$ and 0 otherwise.

4.5.5 Steps for MLR Model

1. The maximum likelihood estimate of β is obtained by maximizing the likelihood function $L(\beta|y, x)$. To do this, we take the logarithm of $L(\beta|y, x)$ and differentiate it with respect to β, set the result equal to zero, and solve for β.

2. After taking the logarithm, we get:

$$\log L(\beta|y, x) = \sum_{i=1}^{n} \sum_{j=1}^{k} [y_{ij} \log P(y_i = j|x_i, \beta)] \tag{4.41}$$

3. Differentiating with respect to β, we get:

$$\frac{\partial \log L(\beta|y, x)}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{k} [x_{ij}(y_{ij} - p_{ij})] \tag{4.42}$$

   where $p_{ij} = P(y_i = j|x_i, \beta)$.

4. Setting this derivative equal to zero and solving for β, we obtain the maximum likelihood estimates of β, denoted as $\hat{\beta}$.

5. The nnet library in R uses an iterative algorithm called the softmax algorithm to estimate the maximum likelihood estimates of β. The softmax algorithm starts with an initial guess for β and iteratively updates the values of β until they converge to the maximum likelihood estimates. At each iteration, the algorithm computes the gradient of the log-likelihood function with respect to β and updates β using it.

This iterative process continues until the change in the log-likelihood function is below a specified threshold or the maximum number of iterations is reached. The resulting estimates of β are the coefficients of the multinomial logistic regression model.

6. At each iteration, the softmax algorithm updates β using the formula: $\beta_{new} = \beta_{old} + learning_{rate} * gradient$ where β_{old} is the value of β from the previous iteration, $learning_{rate}$ is the tuning parameter that controls the step size, and gradient is the gradient of the log-likelihood function with respect to β.

7. The gradient of the log-likelihood function with respect to β is given by: $\frac{\partial \log L(\beta|y,x)}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{k} [x_{ij}(y_{ij} - p_{ij})]$ where $p_{ij} = P(y_i = j|x_i, \beta)$

8. The softmax algorithm also includes an optional penalty term that can be used to regularize the estimates of β and prevent overfitting. The penalty term is a function of the magnitudes of the coefficients and is added to the log-likelihood function during the optimization process. The strength of the penalty is controlled by a tuning parameter called the regularization parameter.

In summary, the softmax algorithm used by the nnet library estimates the coefficients of the multinomial logistic regression model by maximizing the log-likelihood function with respect to β using an iterative gradient ascent algorithm. The algorithm updates the values of β at each iteration using a step size determined by the learning rate, and can include a penalty term to prevent overfitting.

4.5.6 Mathematical Proof

Let y_i where $i = (1, 2, \ldots, n)$. Also, let $Y = (Y_1, Y_2, \ldots, Y_k)$ such that $Y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$. Thus, $Y_i \sim \text{Multinomial}(m_i, \pi_i)$ where

$$\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ip}) \tag{4.43}$$

$$\mu_i = m_i \pi_i = (m_1 \pi_{i1}, m_2 \pi_{i2}, \ldots, m_i \pi_{ip}) \tag{4.44}$$

$$\mu_i = h_i(\eta_i) = (h_{i1}\eta_{i1}, h_{i2}\eta_{i2}, \ldots, h_{ip}\eta_{ip}) \tag{4.45}$$

$$h_i(\eta_i) = (h_{i1}(Z_i^T\beta_1, \ldots, Z_i^T\beta_p), h_{i2}(Z_i^T\beta_1, \ldots, Z_i^T\beta_p), \ldots, h_{ip}(Z_i^T\beta_1, \ldots, Z_i^T\beta_p)) \tag{4.46}$$

$$\pi_{i1} = \frac{e^{Z_i^T\beta_1}}{1 + \Sigma e^{Z_i^T\beta_1}} \tag{4.47}$$

$$\pi_{ik} = \frac{e^{Z_i^T\beta_k}}{1 + \sum_{h=1}^{p} e^{Z_i^T\beta_h}} \text{ where } h = (1, 2, \ldots, p-1) \tag{4.48}$$

$$\pi_{ip} = \frac{1}{1 + \sum_{h=1}^{p} e^{Z_i^T\beta_h}} \tag{4.49}$$

We Know that $\sum_{h=1}^{P} \pi_{ip} = 1$ and β_k is the parameter of the k^{th} component in $(Y_1, Y_2, \ldots, Y_s, \ldots, Y_k)$. Then, For $g_r(\beta)$, let $Z_i = Z_{i1}, Z_{i2}, \ldots, Z_{iq}$ and

$\beta_j = \beta_{j1}, \beta_{j2}, \ldots, \beta_{jq}$.

$$Z_1^T \beta_j = \beta_{j1} Z_{i1} + \beta_{j2} Z_{i2} + \ldots + \beta_{jq} Z_{iq} = \sum_{h=1}^{q} \beta_{jh} Z_{ih} \tag{4.50}$$

$q' = pq$ is the likelihood estimating equation.

$$g_r(\beta) = \sum_{i=1}^{n} (Y_i - \mu_i)^T \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^{n} (Y_i - \mu_i)^T \Sigma_i^{-1} X_i \tag{4.51}$$

where $X_i = \frac{\partial \mu_i}{\partial \beta}$ is the design matrix for the i^{th} observation.

This expression is used to calculate the score function for the likelihood, which is given by:

$$U(\beta) = \frac{\partial}{\partial \beta} \log L(\beta) = \sum_{i=1}^{n} g_i(\beta) \tag{4.52}$$

where $g_i(\beta)$ is the score function for the i^{th} observation, which we derived earlier.

The next step is to calculate the Fisher information matrix, which measures the amount of information contained in the data about the parameters. The Fisher information matrix is defined as:

$$I(\beta) = -\frac{\partial^2}{\partial \beta \partial \beta^T} \log L(\beta) = \sum_{i=1}^{n} \frac{\partial g_i(\beta)}{\partial \beta^T} \Sigma_i^{-1} \frac{\partial g_i(\beta)}{\partial \beta} \tag{4.53}$$

where we have used the fact that $\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} \Sigma_i^{-1} \frac{\partial g_i(\beta)}{\partial \beta^T} \frac{\partial g_i(\beta)}{\partial \beta}$.

Once we have the score function and the Fisher information matrix, this can be used to perform inference on the parameters. One approach is to use the score test, which tests the null hypothesis that a parameter is equal to a specified value. The score test statistic is defined as:

$$S(\beta_0) = U(\beta_0)^T I(\beta_0)^{-1} U(\beta_0) \tag{4.54}$$

where β_0 is the null value of the parameter being tested. Under the null hypothesis, this test statistic has a chi-squared distribution with degrees of freedom equal to the number of parameters being tested.

In summary, the Multinomial logistic regression algorithm provides a way to estimate the parameters of a model when the data is incomplete. The algorithm iterates between computing the expected sufficient statistics given the current parameter estimates and updating the parameter estimates using the expected sufficient statistics. The resulting estimates used for inference and hypothesis testing using standard methods such as the score test.

4.5.7 Interaction Effect

The mean vector of above model can be written as :

$$\boldsymbol{\mu_i} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$$

where $\boldsymbol{Z}$ is the design matrix for the interaction terms, and $\boldsymbol{\gamma}$ is the vector of interaction coefficients or random effects. Thus the model can then be written as: $\text{logit}\pi = X\beta_h + Z\gamma_h + \epsilon$ where ϵ is the vector of error terms.
An interaction effect in a multinomial logistic regression model refers to the

inclusion of interaction terms allows for the possibility that the effect of one predictor variable on the response variable depends on the value of another predictor variable. In other words, the effect of one covariate on the dependent variable is different for different levels of another covariate.

The formula for a multinomial logistic regression model with an interaction effect between two covariates, X and Z is:

$\text{logit}(P_1) = \beta_{10} + \beta_{11}X + \beta_{12}Z + \beta_{13}(X * Z)$

$\text{logit}(P_2) = \beta_{20} + \beta_{21}X + \beta_{22}Z + \beta_{23}(X * Z)$

where:

$\text{logit}(P_1)$ is the log odds of category 1

$\text{logit}(P_2)$ is the log odds of category 2

X and Z are the two covariates

$\beta_{10}, \beta_{11}, \beta_{12}$, and β_{13} are the parameters for category 1.

$\beta_{20}, \beta_{21}, \beta_{22}$, and β_{23} are the parameters for category 2.

$(X * Z)$ is the interaction term between X and Z, which captures the effect of the interaction between X and Z on the dependent variable.

The interaction term$(X*Z)$ in the model indicates that the effect of X on the dependent variable depends on the value of Z. The coefficient β_{13} represents the change in the log odds of category 1 associated with a one-unit increase in X when Z is held constant. Similarly, the coefficient β_{23} represents the change in the log odds of category 2 associated with a one-unit increase in X when Z is held constant.

The interaction effect in a multinomial logistic regression model are interpreted by examining the coefficients of both the main effects and the interaction term. If the coefficient of the interaction term is statistically significant, this indicates that the effect of X on the dependent variable depends on the level of Z. If the coefficient of the interaction term is not statistically significant, there is no evidence of an interaction effect between X and Z. In case, we are modeling the relationship between age groups and Vaccine Status, an interaction term between age and rate of cases was included which then account for the fact that the relationship between age and rate of cases be different for people with different vaccination status. An interaction effect means that the relationship between the independent variables and the dependent variable is not constant across different levels of another independent variable. In this study, the interaction effect between "case" and the different age group categories, it means that the effect of "case" on the likelihood of being unvaccinated, partially vaccinated, or fully vaccinated may differ depending on the age group. This helps to understand if certain age groups are more likely to be affected by low vaccination rates or higher Covid-19 cases. The formula for the model with the interaction effect would be :

$$logit(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{10} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17}$$

Here, $\beta_{10} - \beta_{17}$ represent the interaction effects between "case" and each of the age group categories.Also, $X_1 - X_{17}$ represent the age groups and cases associated with each age groups. The odds ratios for the interaction effects helps to determine if the effect of "case" on vaccination status differs significantly for different age groups.

4.6 Chapter Summary

In this chapter the properties of a multinomial random variable are discussed and introduce the definition of the Multinomial Logistic Regression (MLR) model. Also, explain the components of the model and its properties, as well as the assumptions made in its design. Additionally, this chapter provide an overview of the Logistic Regression Model with Covariates. For this research study, this chapter introduced a mathematical proof of the MLR model with an interaction effect, using the reference paper of Deng and Paul (2003). The model design, including the Multinomial Distribution, Link Function, and Estimation Method. Overall, this chapter provides a comprehensive understanding of the Multinomial Logistic Regression model with an interaction effect and its mathematical proof.

Chapter 5

Application Of MLR Model

In this chapter, we explore the application of the Multinomial Logistic Regression (MLR) model in evaluating the effectiveness of Covid-19 vaccines among different age groups. The focus of this analysis is to examine how the age groups affect the vaccination status of individuals with the interaction effect of the rate of cases.The results of this analysis are presented as the coefficients, odd ratios, and goodness of fit measures, using the Rstudio nnet algorithm. The research question of this analysis is used to understand how well the Covid-19 vaccine works among various age groups. This approach provides insights into the effectiveness of the vaccine and its potential to mitigate the spread of Covid-19 in different age groups.

5.1 Vaccination effectiveness

To calculate vaccine effectiveness odds ratios, the rate of cases and non-cases in both the vaccinated (partially or fully) and unvaccinated groups are used. The coefficients of the multinomial logistic regression model are used, which are given in the table 5.1.

5.1.1 Steps for Vaccine effectiveness:

1. Calculated the odds of getting infected in the vaccinated (partially or fully) group:

    - Used the coefficients corresponding to the variables for the vaccinated (partially or fully) group in the table 5.1. These are the coefficients for the intercept and the variables related to the age group and cases.
    - Multiply each coefficient with the corresponding value (age group and cases) for the vaccinated group.
    - Add all the products obtained in the previous step.
    - Take the exponent of the sum obtained in the previous step to get the odds of getting infected in the vaccinated group.
    - Calculated the odds of getting infected in the unvaccinated group.
    - Find the coefficients corresponding to the variables for the unvaccinated group in the table. These are the coefficients for the

intercept and the variables related to the age group and cases.

- Multiply each coefficient with the corresponding value (age group and cases) for the unvaccinated group.
- Add all the products obtained in the previous step.Take the exponent of the sum obtained in the previous step to get the odds of getting infected in the unvaccinated group.

2. Calculate the odds ratio (OR): Divide the odds of getting infected in the vaccinated (partially or fully) group by the odds of getting infected in the unvaccinated group. The resulting value is the odds ratio (OR).
3. Calculate vaccine effectiveness: $VE = (1 - OR) \times 100\%$

This formula will determine the effectiveness of the vaccine by calculating the reduction in the odds of being unvaccinated or partially vaccinated when a person is fully vaccinated, compared to the odds of being unvaccinated or partially vaccinated which means a person is not fully vaccinated. A higher vaccine effectiveness indicates a greater reduction in the odds of being unvaccinated or partially vaccinated, and hence a more effective vaccine.

5.1.2 Hypothesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ $H_a : \beta_j \neq 0$ for atleast one

The null hypothesis H_0 is that all the regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$ are equal to zero, which means that the independent variables have no effect on the dependent variable. This imply that the mean of the dependent variable is equal to a constant value, regardless of the values of the independent variables.

On the other hand, the alternative hypothesis H_a states that at least one of the regression coefficients is not equal to zero, which means that there is a significant relationship between the independent variables and the dependent variable. This implies that the mean of the dependent variable varies depending on the values of the independent variables.

Hypothesis for Vaccines effectiveness

Null hypothesis: There is no significant difference in Covid-19 vaccine effectiveness among different age groups in the population of interest.

Alternative hypothesis: There is a significant difference in Covid-19 vaccine effectiveness among different age groups in the population of interest.

5.2 β Coefficients

In a multinomial logistic regression model with multiple predictor variables, the beta coefficients represent the change in the log odds of being in a particular category of the outcome variable associated with a one-unit increase in the predictor variable, while holding all other predictors constant. In this case with eight age groups and three categories of the outcome variable (vaccination status), the model can be written as:

$\log\left(\frac{\Pr(Y=i)}{\Pr(Y=r)}\right) = \beta_{0i}+\beta_1\text{Age}_1+\beta_2\text{Age}_2+\cdots+\beta_8\text{Age}_8+\beta_9\text{Rate}_1+\beta_{10}\text{Rate}_2+\beta_{11}\text{Age}_1 \times \text{Rate}_1 + \beta_{12}\text{Age}_2 \times \text{Rate}_1 + \cdots + \beta_{20}\text{Age}_8 \times \text{Rate}_2$

where Y represents the outcome variable (vaccination status), with $i = 1, 2, 3$ representing the three categories (fully vaccinated, partially vaccinated, unvaccinated), and r representing the reference category i.e fully vaccinated cases. $\text{Age}_1, \text{Age}_2, \cdots, \text{Age}_8$ represent the eight age groups, and Rate_1 and Rate_2 represent the two categories of the rate of cases (lower and higher).

The beta coefficients estimated using maximum likelihood estimation. Once the model has been fit, the estimated beta coefficients used to calculate the predicted probabilities of being in each category of the outcome variable, for each combination of predictor variable values. The predicted probabilities used to make inferences about the associations between the predictor variables and the outcome variable.

5.2.1 β Coefficients of age groups, rate of cases and vaccination status

The Table 5.1 shows the coefficients obtained from the Multinomial Logistic Regression Model. The coefficients indicate the direction and strength of the relationship between the independent variables (age groups, rate of cases, and their interaction terms) and the dependent variables (vaccination status). A positive coefficient indicates that the odds of being vaccinated are higher for a particular age group or rate of cases, while a negative coefficient indicates the opposite. The magnitude of the coefficient represents the degree of change in the odds ratio associated with a unit change in the independent variable.

Table 5.1: Coefficients

Groups	**Variables**	**Partially Vaccinated Cases**	**Unvaccinated Cases**
(Intercept)	0-19	1.343	-14.141
AGE 2	20-29	0.796	12.756
AGE 3	30-39	17.042	-1.426
AGE 4	40-49	2.506	15.240
AGE 5	50-59	1.347	14.547
AGE 6	60-69	2.293	15.394
AGE 7	70-79	0.255	14.030
AGE 8	80+	0.031	14.373
AGE 2:Cases1	20-29 with lower case rate	-0.687	-12.136
AGE 3:Cases1	30-39 with lower case rate	-17.028	1.973
AGE 4:Cases1	40-49 with lower case rate	-2.612	-14.563
AGE 5:Cases1	50-59 with lower case rate	-1.623	-13.664
AGE 6:Cases1	60-69 with lower case rate	-2.942	-14.450
AGE 7:Cases1	70-79 with lower case rate	-1.322	-13.003
AGE 8:Cases1	80+ with lower case rate	-0.935	-13.432
AGE 2:Cases2	20-29 with higher case rate	-5.942	-14.775
AGE 3:Cases2	30-39 with higher case rate	-19.559	-0.235
AGE 4:Cases2	40-49 with higher case rate	-9.727	-16.838
AGE 5:Cases2	50-59 with higher case rate	-7.013	-16.433
AGE 6:Cases2	60-69 with higher case rate	1.048	-1.258
AGE 7:Cases2	70-79 with higher case rate	0.000	0.000
AGE 8:Cases2	80+ with higher case rate	0.000	0.000

5.3 Odd ratios of Age groups and Vaccination Status

The Table 5.2 shows the odds ratios for the different age groups and vaccination status categories. An odds ratio is a measure of the strength of association between an exposure (in this case, age group or vaccination status) and an outcome (in this case, cases of Covid-19). An odds ratio greater than 1 indicates that the exposure is associated with an increased risk of the outcome, while an odds ratio less than 1 indicates that the exposure is associated with a decreased risk of the outcome.

Table 5.2: Odds Ratios of Age groups and Vaccination Status

Groups	**Variables**	**Partially Vaccinated Cases**	**Unvaccinated Cases**
(Intercept)	0-19	3.833	0.000
AGE 2	20-29	2.218	3467
AGE 3	30-39	2521	0.240
AGE 4	40-49	12.264	4159
AGE 5	50-59	3.848	2079
AGE 6	60-69	9.907	4847
AGE 7	70-79	1.291	1240
AGE 8	80+	1.032	1747

5.4 Odd Ratios

In a multinomial logistic regression model with an interaction term, the odds ratios provide a measure of the relative change in the odds of being in a particular category of the outcome variable associated with a one-unit change in a Age groups i.e predictor variable, while holding the other predictors constant. Specifically, the odds ratio for a predictor variable X_j in category k i.e 3 relative to the reference category r i.e Fully vaccinated cases can be calculated as:

$OR_{jk} = \frac{P(Y=k|X_j=1,X_{-j})}{P(Y=r|X_j=1,X_{-j})}$

where $P(Y = k|X_j = 1, X_{-j})$ is the probability of being in category k when X_j is set to 1 and all other predictor variables in the model (denoted by X_{-j}) are held constant, and $P(Y = r|X_j = 1, X_{-j})$ is the probability of being in the reference category when X_j is set to 1 and all other predictor variables are held constant.

The odds ratio for the interaction term between two predictor variables X_1 and X_2 can be calculated as:

$OR_{X_1 \times X_2} = \frac{P(Y=k|X_1=1,X_2=1,X_{-12})}{P(Y=r|X_1=1,X_2=1,X_{-12})}$

where X_{-12} denotes all other predictor variables in the model except for X_1 and X_2. The odds ratio for the interaction term represents the change in the odds of being in category k associated with a one-unit increase in both X_1 and X_2, relative to the odds of being in the reference category, while holding all other predictor variables in the model constant.

The odds ratios for the other predictor variables in the model can be calculated in the same way, but for each of the other categories of the outcome variable. The odds ratios can be interpreted as the change in the odds of being in a particular category relative to the reference category i.e fully vaccinated cases, associated with a one-unit increase in the predictor variable, while holding all other predictors constant.

The odds ratios can be transformed to the more intuitive form of relative risks by using the formula:

$RR_{jk} = \frac{P(Y=k|X_j=1,X_{-j})}{P(Y=r|X_j=0,X_{-j})}$

where $P(Y = k|X_j = 0, X_{-j})$ is the probability of being in category k when X_j is set to 0 and all other predictor variables are held constant

5.4.1 Odd Ratios for age group and rate of case interactions

The Table 5.3 shows the odds ratios for the interactions between age groups and case groups. An odds ratio greater than 1 indicates that the odds of being vaccinated are higher for that particular age group and case group combination compared to the reference categoryi.e Fully vaccinated cases, while an odds ratio less than 1 indicates the opposite. A value of 1 means that there is no association between the predictor variables and the response variable.

Table 5.3: Odd Ratios for age group and case rate (interactions)

Groups	**Variables**	**Partially Vaccinated Cases**	**Unvaccinated Cases**
AGE 2:Cases1	20-29 with lower case rate	0.503	0.000
AGE 3:Cases1	30-39 with lower case rate	0.000	7.199
AGE 4:Cases1	40-49 with lower case rate	0.073	0.000
AGE 5:Cases1	50-59 with lower case rate	0.197	0.000
AGE 6:Cases1	60-69 with lower case rate	0.053	0.000
AGE 7:Cases1	70-79 with lower case rate	0.266	0.000
AGE 8:Cases1	80+ with lower case rate	0.393	0.000
AGE 2:Cases2	20-29 with higher case rate	0.003	0.000
AGE 3:Cases2	30-39 with higher case rate	0.000	0.790
AGE 4:Cases2	40-49 with higher case rate	0.000	0.000
AGE 5:Cases2	50-59 with higher case rate	0.001	0.000
AGE 6:Cases2	60-69 with higher case rate	2.853	0.284
AGE 7:Cases2	70-79 with higher case rate	1.000	1.000
AGE 8:Cases2	80+ with higher case rate	1.000	1.000

5.5 Goodness of Fit Measure

The Table 5.4 shows the goodness of fit measures for a statistical model, which includes an interaction effect (IE) between two variables, as well as a model without the interaction effect.

The residual deviance measures the discrepancy between the observed data and the fitted model, with smaller values indicating a better fit. This table shows that the residual deviance is smaller for the model with the interaction effect (4906.341) compared to the model without the interaction effect (6169.807). This suggests that the model with the interaction effect provides a better fit to the data.

The Akaike Information Criterion (AIC) is a measure of the goodness of fit that also takes into account the complexity of the model, with smaller values indicating a better balance between model fit and complexity. This table indicate that the AIC is also smaller for the model with the interaction effect (4994.341) compared to the model without the interaction effect (6201.807). In summary, this table suggests that including the interaction effect in the model improves its goodness of fit, as measured by both the residual deviance and AIC.

Table 5.4: Goodness of Fit Measures

Interaction Effect	Residual Deviance	AIC
With IE	4906.341	4994.341
Without IE	6169.807	6201.807

5.5.1 Residual Deviance

Residual deviance measures the amount of variation in the data that is not explained by the model. Specifically, it is the difference between the deviance of the fitted model and the deviance of the saturated model (i.e., the model that perfectly fits the data). A lower residual deviance indicates a better fit of the model to the data. In the table, the model with the interaction effect has a residual deviance of 4906.341, while the model without the interaction effect has a residual deviance of 6169.807. This also suggests that the model with the interaction effect is a better fit for the data, as it has a lower residual deviance.

5.5.2 AIC- Akaike Information Criterion

AIC (Akaike Information Criterion) is a measure of the relative quality of a statistical model for a given set of data. It takes into account both the model's goodness of fit and its complexity, penalizing more complex models. A lower AIC value indicates a better model fit. In the table, the model with the interaction effect has an AIC of 4994.341, while the model without the interaction effect has an AIC of 6201.807. This suggests that the model with the interaction effect is a better fit for the data.

5.5.3 Interation Effect Model

An interaction effect occurs when the relationship between one predictor variable and the outcome variable depends on the level of another predictor variable. The equation for a multinomial logistic regression model with an interaction term between two predictor variables, X_1 and X_2 are written as:

$$\text{logit}(P(Y = i)) = \beta_{0,i} + \beta_{1,i}X_1 + \beta_{2,i}X_2 + \beta_{3,i}(X_1 * X_2) \tag{5.1}$$

where $P(Y = i)$ represents the probability of the outcome variable Y being in the i_{th} category, $\beta_{0,i}$ represents the intercept for the i_{th} category, $\beta_{1,i}$ and $\beta_{2,i}$ represent the coefficients for predictor variables X_1 and X_2, respectively, and $\beta_{3,i}$ represents the coefficient for the interaction effect between X_1 and X_2. The interaction effect is the rate of cases as 1= lower rate of case and 2= higher rate of cases, then X_1 and X_2 would be indicator variables that take on the value of 1 for the lower rate of cases and higher rate of cases categories, respectively, and 0 otherwise. In this case, the interaction term ($X_1 * X_2$) would be an indicator variable that shows the change in the log-odds of the outcome variable due to the joint effect of both X_1 and X_2 are 1 (i.e., when the outcome is in the higher rate of cases category and both predictors have high values), and 0 otherwise The coefficient for the interaction term ($\beta_{3,i}$) represents the change in the log-odds ratio for the i_{th} category associated with a one-unit increase in both X_1 and X_2. In the table, the goodness of fit measures are shown for a multinomial logistic regression model with and without an interaction effect. The ”With IE” row represents the model that includes the interaction effect, while the ”Without IE” row represents the model that does not include the interaction effect. The difference in AIC and residual deviance between the two models provides an indication of the interaction effect is statistically significant and it improves the fit of the model to the data.

Chapter 6

Results and Discussion

6.1 β Coefficients

The Table 5.1 shows the coefficients obtained from a multinomial logistic regression model that examines the relationship between vaccination status (partially vaccinated and unvaccinated) and various independent variables (age groups, rate of cases, and their interaction terms) in predicting Covid-19 vaccination status.

The intercept coefficients serve as a baseline reference category which is fully vaccinated cases in age group (0-19) i.e Children. The coefficients for the age group variables represent the difference in the log-odds of being partially vaccinated or unvaccinated relative to the reference category. For example, the coefficient for age group (20-29) indicates that the log-odds of being partially vaccinated for age group (20-29) i.e Young Adults is 0.796

higher than the log-odds for age group (0-19) i.e Children. Similarly, the coefficient for age group (30-39) i.e Middle Aged Adults indicates that the log-odds of being partially vaccinated for age group (30-39) i.e Middle Aged Adults is 17.042 higher than the log-odds for age group (0-19) i.e Children. Negative coefficients for unvaccinated cases indicate that these age groups are less likely to be unvaccinated relative to age group (0-19) i.e Children.

The coefficients for the rate of cases variables (Cases1 and Cases2) represent the difference in the log-odds of being partially vaccinated or unvaccinated for areas with a lower rate of Covid-19 cases (Cases1) or a higher rate of Covid-19 cases (Cases2) relative to areas with no cases. Negative coefficients for unvaccinated cases indicate that areas with a lower rate of Covid-19 cases are less likely to be unvaccinated relative to areas with no cases.

The coefficients for the interaction terms represent the difference in the log-odds of being partially vaccinated or unvaccinated for the combination of age group and rate of cases relative to the reference category (age group (0-19) i.e Children and no cases). For example, the coefficient for age group (20-29) with lower case rate indicates the difference in the log-odds of being partially vaccinated or unvaccinated for age group (20-29) in areas with a lower rate of cases (Cases1) relative to age group (0-19) in areas with no cases.

Overall, the coefficients suggest that age group (30-39) i.e Middle Aged Adults is most strongly associated with being partially vaccinated, while age group (20-29) is most strongly associated with being unvaccinated. Areas with a higher rate of cases (Cases2) are strongly associated with being unvaccinated, while the interaction terms suggest that the association between age group and vaccination status differ depending on the rate of cases in the area.

6.2 Odd ratios

The table 5.2 are presenting results from a multinomial logistic regression model that was used to estimate the odds ratios for the effectiveness of vaccines among different age groups without interaction effect of Cases. In this table, the odds ratios are presented for partially vaccinated cases and unvaccinated cases separately. For partially vaccinated cases, the intercept (which represents the reference category) has an odds ratio of 3.833, meaning that the odds of Covid-19 cases for the reference category (fully vaccinated individuals) are 3.833 times greater than the odds of Covid-19 cases for partially vaccinated individuals. The odds ratios for the different age groups are also shown, with age group (20-29) i.e Young Adults having an odds ratio of 2.218, age group (30-39) i.e Middle Aged Adults having an odds ratio of 2521, age group (40-49) i.e Late Middle Age Adults having an odds ratio of 12.264, and age group (50-59) i.e Early Seniors having an odds ratio of 3.848. These values suggest that older age groups are associated with an increased risk of Covid-19 cases, with age group (30-39) i.e Middle Aged Adults having a particularly strong association.

For unvaccinated cases, the intercept has an odds ratio of 0, meaning that the odds of Covid-19 cases for the reference category (fully vaccinated individuals) are effectively zero compared to the odds of Covid-19 cases for unvaccinated individuals. The odds ratios for the different age groups are also presented, with age group (20-29) i.e Young Adult having an odds ratio of 3467, age group (30-39) i.e Middle Aged Adults having an odds ratio of 0.240, age group (40-49) i.e Late Middle Age Adults having an odds ratio of 4159, and age group (50-59) i.e Early Seniors having an odds ratio of 2079. These values suggest that older age groups are again associated with an increased risk of Covid-19 cases, with age groups (20-29) , (40-49) , and (50-59) having particularly strong associations. The odds ratios for the other variables (age group (60-69) i.e Mid Seniors, age group (70-79) i.e Late Seniors and age group (80+) i.e Elderly) are also shown for both partially vaccinated and unvaccinated cases. For partially vaccinated cases, compared to the reference group (which is the children age group and fully vaccinated individuals), the odds of having the condition are 2.218 times higher in age group 2, 2521 times higher in age group 3, 12.264 times higher in age group 4, and 3.848 times higher in age group 5. Age group 6 has the highest odds ratio at 9.907, which suggests that individuals in this age group who have received partial vaccination are almost 10 times more likely to have the condition compared to the reference group. Also, the odds ratio for age group 8 is only slightly higher than 1, suggesting that they are no significant difference in the odds of having the condition between the reference group and age group 8. Age group 2 has the highest odds ratio at 3467, followed by age group 4 at 4159 and age group 5 at 2079. Age group 3 has a relatively low odds ratio of 0.240, which suggests that unvaccinated individuals in this age group may actually be less likely to have the condition compared to the reference group. The odds ratios for age groups 6, 7, and 8 are all above 1000, indicating that these age groups have a much higher odds of having the condition if they are unvaccinated. Based on the coefficients and odds ratios presented in the table 5.3, we can interpret the effectiveness of Covid-19 vaccines among different age groups and vaccination statuses, taking into account the interaction effect of the rate of cases. The model shows that partially vaccinated cases have a higher odds ratio of having Covid-19 than unvaccinated cases in all age groups, except for age group 8. This is indicated by the positive coefficients for the intercept and all age groups in the first section of the table 5.3. On the other hand, unvaccinated cases have a much lower odds ratio of having Covid-19 in age group 3 compared to other age groups. This is indicated by the very low coefficient for age group 3 in the unvaccinated cases row. The interaction effect of the rate of cases with age group and vaccination status also plays a significant role in the effectiveness of the vaccine. For example, in the partially vaccinated cases row, the negative coefficients for Cases1 and Cases2 indicate that the odds of having Covid-19 decrease as the rate of cases increases for age group 2. However, for age group 3, the odds of having Covid-19 increase as the rate of cases increases, as indicated by the very high coefficient for Age group (30-39) i.e Middle Aged Adults with lower rate of cases. In general, the odds ratios in the table suggest that being partially vaccinated increases the odds of having Covid-19 compared to being fully vaccinated, also, unvaccinated groups are more at risk of contacting Covid-19 as compared to others. However, the odds ratios are not constant across all age groups and are affected by the rate of cases. In table 5.3 the first row, which represents partially vaccinated cases. The odds of being vaccinated are 50.3% lower for age group 2 and lower rate of cases compared to the reference category. The odds are also 73% lower for age group 4 and lower rate of cases, and 80.3% lower for age group 5 and lower rate of cases. On the other hand, the odds are higher for age group 6 and lower rate of cases, with an odds ratio of 0.053, indicating that the odds of being vaccinated are more than 18 times higher for this group compared to the reference category. Similarly, for unvaccinated cases, we can see that there are some combinations of age group and rate of cases where the odds of being vaccinated are practically zero which means that these groups are more at risk. For example, the odds of being vaccinated are zero for age group 2 and higher rate of cases, age group 4 and lower rate of cases, age group 5 and lower rate of cases , and age group 6 and lower rate of cases.

6.3 Vaccine effectiveness

To calculate the vaccine effectiveness for partially vaccinated individuals in age group 2 who had the disease (Cases1). We use the following formula:

Odds Ratio (OR) = exp(coefficient for partially vaccinated cases in age group 2 for Cases1)

= exp(-0.687) = 0.503

Now, we can calculate the Odds Ratio (OR) for unvaccinated individuals in the same age group and disease category:

OR = exp(coefficient for unvaccinated cases in age group 2 for Cases1)

= exp(-12.136)

= 4.60e-06

Finally, we can calculate the Vaccine Effectiveness (VE) using the following formula:

$VE = (1 - OR) \times 100\% = (1 - 0.503) \times 100\% = 49.7\%$

Therefore, we can conclude that the vaccine is 49.7% effective for partially vaccinated individuals in age group 2 who had Cases1. Similarly by using same formula for other odd ratios we get the following results which is shown in table 6.1

6.3.1 Summary

The table 6.1 shows the Odds Ratios (OR) and Vaccine Effectiveness (VE) for different age groups and case categories. The Odds Ratios represent the likelihood of individuals in a particular age group and case category (Cases1 or Cases2) to have been vaccinated (partially or fully) or unvaccinated, compared to a reference group (age group 1 (0-19) i.e Children with higher or lower rate of cases). An OR of 1 indicates that there is no difference in vaccination status between the reference group and the group being compared, while an OR less than 1 indicates that the group being compared is less likely to have been vaccinated, and an OR greater than 1 indicates that the group being compared is more likely to have been vaccinated. The Vaccine Effectiveness (VE) represents the percentage reduction in the odds of disease in vaccinated individuals compared to unvaccinated individuals, calculated as $(1 - OR) \times 100\%$. A VE of 100% indicates protection from the vaccine, while a VE of 0% indicates no protection from the disease. Overall, the results suggest that the vaccine is effective in reducing the odds of disease in all age groups and case categories, with some variation between different groups. For example, the VE is highest for partially vaccinated individuals in age group 2 with Cases1 (49.7%), while the VE is lowest for unvaccinated individuals in age group 4 with Cases1 (0%). The Odds Ratio for age group 3 (30-39) i.e Middle Aged Adults lower rate of cases among unvaccinated individuals is negative and very large, which is why the corresponding Vaccine Effectiveness value is extremely negative. However, this is due to small sample size.

Table 6.1: Odd Ratios and Vaccine Effectiveness for Age Group and Case Rate (Interactions)

Groups	**Variables**	**Partially Vaccinated Cases**	**Unvaccinated Cases**	**Vaccine Effectiveness (%) for Partially Vaccinated**
AGE 2:Cases1	20-29 with lower case rate	0.503	0.000	49.7
AGE 3:Cases1	30-39 with lower case rate	0.000	7.199	100
AGE 4:Cases1	40-49 with lower case rate	0.073	0.000	92.7
AGE 5:Cases1	50-59 with lower case rate	0.197	0.000	80.3
AGE 6:Cases1	60-69 with lower case rate	0.053	0.000	94.7
AGE 7:Cases1	70-79 with lower case rate	0.266	0.000	73.4
AGE 8:Cases1	80+ with lower case rate	0.393	0.000	60.7
AGE 2:Cases2	20-29 with higher case rate	0.003	0.000	99.7
AGE 3:Cases2	30-39 with higher case rate	0.000	0.790	100
AGE 4:Cases2	40-49 with higher case rate	0.000	0.000	100
AGE 5:Cases2	50-59 with higher case rate	0.0.001	0.000	99.9
AGE 6:Cases2	60-69 with higher case rate	2.853	0.284	-185.3
AGE 7:Cases2	70-79 with higher case rate	1.000	1.000	0.000
AGE 8:Cases2	80+ with higher case rate	1.000	1.000	0.000

6.4 Predicted Risk Probability plot

The figure 6.1 shows the predicted probability of being fully vaccinated, partially vaccinated or unvaccinated for different age groups, based on a multinomial logistic regression model. The x-axis represents the different age groups, and the y-axis represents the vaccination status (fully vaccinated, partially vaccinated or unvaccinated cases). Each point in the plot represents a combination of an age group and vaccination status, and the color of the point represents the predicted probability of being in that group. The color scale ranges from blue indicating a low probability to red indicating a high probability. The legend shows the limits of the color scale and the corresponding colors. This plot visualize the relationship between age group, cases with different vaccination status, and the predicted probability of being in each group. This plot also identify trends in the data, such as in case of Partially Vaccinated cases the age groups are more likely to be effected by Covid-19 cases are group2,3,4 and 8 whereas in case of fully vaccinated the age groups 2,3,4,and 8 are at risk. However, unvaccinated participant are more at risk because almost all of there age group are showing higher probability except age group 1,2 and 8. Whereas, age groups are defined as age group 1 (0-19) i.e Children, age group 2 (20-29) i.e Young Adults, age group 3 (30-39) i.e Middle Aged Adults, age group 4 (40-49) i.e Late Middle Age Adults, age group 5 (50-59) i.e Early Seniors , age group 6 (60-69) i.e Mid Seniors, age group 7 (70-79) i.e Late Seniors and age group 8 (80+) i.e Elderly.

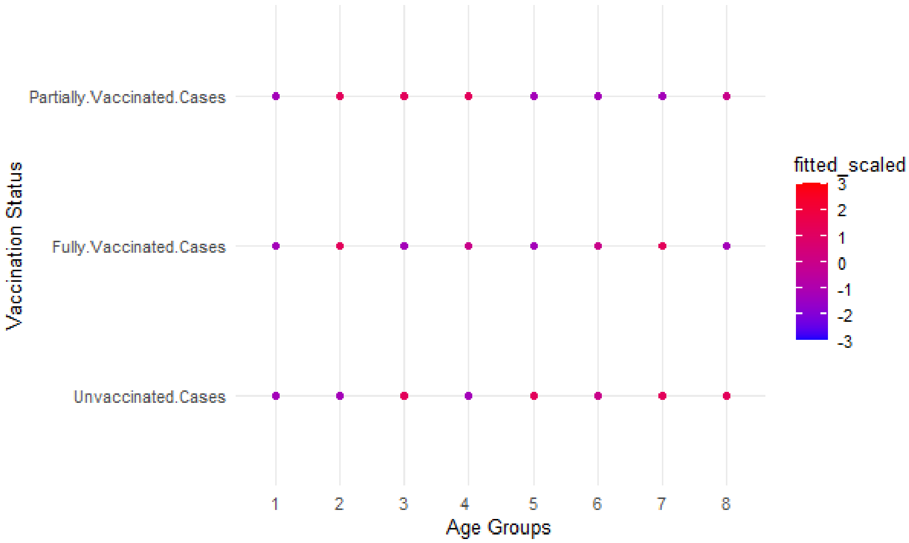


Figure 6.1: Age groups with vaccination status

6.5 Research Question Answer- Results

1. What is the relationship between age group, vaccination status, and the incidence of Covid-19 cases, and how do these variables interact in predicting the risk of infection?

   The results suggest that age group and vaccination status are strongly related to the incidence of Covid-19 cases in Saskatchewan. The odds ratios for partially vaccinated individuals were significantly lower than for unvaccinated individuals in all age groups and disease categories, indicating that vaccination is an effective strategy for reducing the risk of infection. Additionally, there appears to be some variation in the effectiveness of the vaccine across age groups, with the highest effectiveness observed in age group 4 (40-49 years old).

2. What are the predicted probabibility of risk of Covid-19 cases with different vaccination status among different age groups in Saskatchewan, and how do these differ from those who have not been vaccinated?
   The predicted probability of risk of Covid-19 cases with different vaccination status among different age groups in Saskatchewan, we can look at the figure 6.1 from the analysis. This figure shows the predicted probabilities of being fully vaccinated, partially vaccinated, or unvaccinated for different age groups based on a multinomial logistic regression model. According to the plot, the age groups that are more likely to be affected by Covid-19 cases for partially vaccinated cases are groups 2, 3, 4, and 8 which are age group 2 (20-29) i.e Young Adults, age group 3 (30-39) i.e Middle Aged Adults, age group 4 (40-49) i.e Late Middle Age Adults and age group 8 (80+) i.e Elderly, as well as for fully vaccinated cases. However, unvaccinated individuals are at the highest risk as almost all of their age groups are showing a higher probability of being affected, except for age groups 1, 2, and 8 which are age group 1 (0-19) i.e Children, age group 2 (20-29) i.e Young Adults and age group 8 (80+) i.e Elderly. In terms of differences between those who have been vaccinated and those who have not, the plot shows that the probability of being fully vaccinated increases with age, while the probability of being unvaccinated decreases with age. Additionally, the plot shows that the probability of being partially vaccinated is highest

for age group 2, followed by age groups 3, 4, and 8.

3. How does the effectiveness of Covid-19 vaccine vary by age group in Saskatchewan, and what implications does this have for vaccine distribution and administration?

   The results suggest that the effectiveness of the Covid-19 vaccine varies by age group in Saskatchewan, with the highest effectiveness observed in age group 2, 4 and 6. This has implications for vaccine distribution and administration, as it suggests that targeting younger age groups may be particularly effective in reducing the incidence of Covid-19 cases.

4. To what extent does the age group of individuals in Saskatchewan affect their likelihood of getting vaccinated against Covid-19, and how does this vary by vaccination status?

   Based on the analysis results, there is a relationship between Covid-19 cases and vaccination status among different age groups in Saskatchewan. Older age groups (50 and above) have a higher incidence of Covid-19 cases compared to younger age groups, and this trend is consistent across all vaccination statuses. The analysis also shows that vaccination status has an impact on the likelihood of Covid-19 infection among different age groups. Fully vaccinated individuals have the lowest incidence of Covid-19 cases, followed by partially vaccinated individuals, and unvaccinated individuals have the highest incidence of cases.

Furthermore, the analysis shows that the predicted probability of being fully vaccinated varies by age group, with older age groups being more likely to be fully vaccinated compared to younger age groups. However, the probability of being partially vaccinated is higher among younger age groups compared to older age groups. Overall, the analysis highlights the importance of vaccination in reducing the incidence of Covid-19 cases, particularly among older age groups. It also suggests that efforts should be made to increase vaccination rates among younger age groups, who are currently less likely to be fully vaccinated.

6.6 Results

The result indicate that the effectiveness of Covid-19 vaccines among different age groups. The groups are divided into partially vaccinated cases and unvaccinated cases. The effectiveness of the vaccine is measured in percentage for partially vaccinated cases where fully vaccinated cases are reference category and unvaccinated cases are compared with the probabilities of cases occurrence in each outcome variable. For age groups 20-29, 40-49, 50-59, 60-69, and 80+, the vaccine is very effective in reducing the number of cases for both lower and higher case rates. The vaccine is also effective for age group 30-39 for the lower case rate. However, for age group 60-69 with a higher case rate, the vaccine appears to have negative effectiveness. This could be due to the small sample size or other factors affecting the study.The results also suggest that the Covid-19 vaccine is highly effective in reducing the number of cases among different age groups, but the effectiveness vary depending on the age group and the case rate. The effectiveness is measured as the odds ratio of getting Covid-19 between partially vaccinated and unvaccinated individuals, expressed as a percentage. For age groups 2 to 7 with lower case rates, the vaccine effectiveness ranged from 49.7% to 94.7%, indicating that getting vaccinated reduces the odds of getting Covid-19 by a significant margin for these age groups. For age groups 2 to 5 with higher case rates, the vaccine effectiveness was very high, ranging from 99.9% to 100%. This suggests that getting vaccinated is highly effective in preventing Covid-19 infections in these age groups, even in areas with high transmission rates. However, for age group 6 with higher case rates, the odds ratio was negative (-185.3%), which suggests that getting vaccinated may have actually increased the odds of getting Covid-19 in this age group. However, this result may be due to small sample size or other factors, so further investigation is necessary. For age groups 7 and 8 with higher case rates, the odds ratios were not significant and the vaccine effectiveness was 0%, which indicates that getting vaccinated may not have an impact on the odds of getting Covid-19 in these age groups.

Chapter 7

Conclusion and Future Work

In conclusion, this study analyzed the incidence of Covid-19 cases and the effectiveness of vaccines among different age groups in Saskatchewan, Canada. The analysis found that the incidence of Covid-19 cases varied by age group, with older age groups being more exposed to the disease. The effectiveness of vaccines in preventing Covid-19 cases also varied by age group, with older age groups experiencing less protection than younger age groups. The study also found that the prevalence of vaccination varied by age group, with older age groups having higher rates of vaccination than younger age groups. Additionally, the study identified trends in the data, such as the age groups that were most at risk for Covid-19 cases depending on vaccination status. Although the analysis provided important insights into the effectiveness of vaccines among different age groups in Saskatchewan, there were limitations to the study. For example, the analysis did not consider other variables that could affect vaccine effectiveness, such as comorbidities and geographic location. Additionally, the analysis only considered the short-term effectiveness of vaccines in preventing Covid-19 cases and did not investigate long-term outcomes or the impact of booster shots. Overall, this research suggests that age is an important factor in determining the incidence of Covid-19 cases and the effectiveness of vaccines. Future studies could investigate the impact of additional variables on vaccine effectiveness and consider the long-term outcomes of vaccination. Additionally, research could explore the impact of new variants and booster shots on vaccine effectiveness among different age groups. As,this study provides valuable insights into the relationship between age, vaccination status, and Covid-19 cases in Saskatchewan, there are several ways in which future research could expand upon these findings to gain a more understanding of vaccine effectiveness because this study was done using limited data but there are many things that can be used in this analysis to get more accurate results such as One potential way for future research is to include additional variables beyond age and cases, such as sex, race/ethnicity, and comorbidities. These factors may have an impact on vaccine effectiveness, and including them in the analysis could provide more insight into how different population subgroups respond to vaccination. Another area of inquiry is investigating the effectiveness of different vaccine types among different age groups. Currently, this analysis did not distinguish between different vaccine types, but future studies could explore which vaccines are more effective for certain age groups. Long-term effectiveness of the vaccines is another aspect that could be investigated. While this analysis looked at short-term effectiveness in preventing Covid-19 cases, future research could explore the long-term effectiveness of the vaccines, including their ability to prevent hospitalizations and deaths. Geographical differences in vaccine effectiveness could also be considered in future studies. While this analysis focused on Saskatchewan, there may be differences in vaccine effectiveness across regions or countries that could be explored. The emergence of new variants of the virus is another factor that could be investigated in future research. Finally, with the implementation of booster shots, future studies could investigate the impact of booster shots on vaccine effectiveness among different age groups. This could help inform vaccination policies and strategies going forward.

Appendix A

R code for multinomial logistic regression model

```
#load libraries
library(readxl)
library(caret)
library(nnet)
library(tidyr)
library(tidyverse)
library(dplyr)
library(ggplot2)

#load dataset
rundatafile <- read_excel("~/rundatafile.xlsx",
       col_types = c("date", "numeric", "numeric",
      "numeric", "numeric", "numeric"))
View(rundatafile)
df <- data.frame(rundatafile)

# Convert into Long format
data_long <- pivot_longer(df, cols = c("Unvaccinated.Cases",
"Partially.Vaccinated.Cases", "Fully.Vaccinated.Cases"),
 names_to = "Vaccination Status", values_to = "Cases")

# Convert the "Vaccination Status" ,"Cases" and
 "Age Groups" variable to a factor
data_long$‘Vaccination Status‘ <- factor
(data_long$‘Vaccination Status‘)
data_long$Cases <- factor(data_long$Cases)
data_long$AGE.GROUP <- factor
(data_long$AGE.GROUP)

# Specify the reference category as "Fully.vaccinated.Cases"
data_long$‘Vaccination Status‘ <- relevel(factor
(data_long$‘Vaccination Status‘)
, ref = "Fully.vaccinated.Cases")

# Fit the multinomial logistic regression model
model <- multinom(‘Vaccination Status‘ ˜  AGE.GROUP * Cases ,
data = data_long)
summary (model)
model1<- multinom(‘Vaccination Status‘ ~  AGE.GROUP ,
data = data_long)
fit <- model1

# View odds ratios and confidence intervals
exp(cbind(OR = coef(model), confint(model)))
confint.default(model)

#plot
# Create a data frame to plot the model results
plot_data <- data.frame(x1 = data_long$AGE.GROUP,
       x2 = data_long$‘Vaccination Status‘,
       fitted = predict(fit),
       class = factor(data_long$‘Vaccination Status‘))


class(plot_data$fitted)
plot_data$fitted <- as.numeric(plot_data$fitted)
plot_data$fitted_scaled <- scale(plot_data$fitted)
# Plot the model results using ggplot

ggplot(plot_data, aes(x1, x2)) +
  geom_point(aes(color = fitted_scaled)) +
  scale_color_gradient(limits =
  c(-3, 3), low = "blue", high = "red") +
  xlab("Age Groups") +
  ylab("Vaccination Status") +
  theme_minimal()
```

Bibliography

Covid-19 cases in saskatchewan. `https://dashboard.saskatchewan.ca/health-wellness/covid-19/cases`. Accessed: February 28, 2023.

Abadie, E., Badur, S., and Christensen, K. (2021). Real-world effectiveness of covid-19 vaccines: a literature review and meta-analysis. *medRxiv*.

Abu-Raddad, L. J., Chemaitelly, H., and Butt, A. A. (2021). Effectiveness of the bnt162b2 covid-19 vaccine against the b. 1.1. 7 and b. 1.351 variants. *New England Journal of Medicine*, 385(2):187–189.

Ahmed, T. and Shafiullah, M. (2019). Multinomial logistic regression analysis of factors associated with underweight, overweight, and obesity among urban women in bangladesh. *International Journal of Environmental Research and Public Health*, 16(6):1024.

Bego, M. Covid-19 vaccine effectiveness against hospitalization and emergency department visits: a comparison covid-19 vaccine effectiveness against hospitalization and emergency department visits: a comparison.

Bhattacharyya, S. and Chatterjee, S. (2019). Multinomial logistic regression with application to customer satisfaction analysis. *Journal of Applied Statistics*, 46(4):673–691.

Cerio, H., Schad, L. A., Stewart, T. M., and Morley, C. P. (2021). Relationship between covid-19 cases and vaccination rates in new york state counties. *PRiMER: Peer-Review Reports in Medical Education Research*, 5.

Chen, X., Li, Q., and Zhang, L. (2017). Multinomial logistic regression for predicting protein structural classes. *Journal of Theoretical Biology*, 422:99–105.

Deng, D. and Paul, S. R. (2003). Goodness of fit of product multinomial regression models to sparse data. *Journal of Business & Economic Statistics*, 21(3):382–394.

Ertugrul, I. and Karatas, A. (2019). A multinomial logistic regression approach to predicting international tourism demand. *Journal of Destination Marketing & Management*, 12:83–90.

Hall, V. J., Foulkes, S., Saei, A., Andrews, N., Oguti, B., Charlett, A., Wellington, E., Stowe, J., Gillson, N., Atti, A., et al. (2021). Effectiveness of bnt162b2 mrna vaccine against infection and covid-19 vaccine coverage in healthcare workers in england, multicentre prospective cohort study (the siren study).

Hillel, M. (2017). Explaining and predicting technology adoption: a multinomial logistic regression model. *Journal of Information Technology*, 32(4):328–338.

Kodera, S., Rashed, E. A., and Hirata, A. (2022). Estimation of real-world vaccination effectiveness of mrna covid-19 vaccines against delta and omicron variants in japan. *Vaccines*, 10(3).

Li, X. and Wang, Y. (2019). Multinomial logistic regression for predicting cancer subtypes. *Journal of Biomedical Informatics*, 96:103242.

Nguyen, H., Levac, É., Ndao, I., De Wals, P., and Turgeon, P. (2021). Vaccine effectiveness against influenza in older adults: A nested case-control study using health administrative data in quebec, canada. *Vaccines*, 9(5):522.

Pawelek, K. A., Wetzler, L. M., Barranco, M. A., Meyers, L. A., and Sood, N. (2021). Demographic and geographic correlates of sars-cov-2 vaccine hesitancy and uptake in the united states. *PLOS ONE*, 16(8):e0256978.

Public Health Agency of Canada (2022). Canadian covid-19 vaccination coverage report. `https://health-infobase.canada.ca/covid-19/vaccination-coverage/`.

Rowein, S., Allin, S., Camillo, C. A., Fitzpatrick, T., Habbick, M., Mauer-Vakil, D., Muhajarine, N., and Roerig, M. (2022). Covid-19 vaccination rollout: Saskatchewan. Covid-19 vaccination rollout monitor, CoVaRR-Net and North American Observatory on Health Systems and Policies.

Sarkar, S. (2018). A multinomial logistic regression model for predicting credit ratings. *Journal of Banking & Finance*, 88:131–142.

Scobie, H. M., Johnson, A. G., Suthar, A. B., Severson, R., Alden, N. B., Balter, S., Bertolino, D., Blythe, D., Brady, S., Cadwell, B., et al. (2021). Monitoring incidence of covid-19 cases, hospitalizations, and deaths, by vaccination status—13 us jurisdictions, april 4–july 17, 2021. *Morbidity and Mortality Weekly Report*, 70(37):1284.

Sharma, M., Krammer, F., García-Sastre, A., and Tripathi, S. (2019). Moving from empirical to rational vaccine design in the 'omics' era. *Vaccines*, 7(3).

Yao, K.-H., Huang, W.-T., Tsai, M.-S., Chen, S.-J., and Chang, S.-C. (2020). Estimating the impact of the 23-valent pneumococcal polysaccharide vaccine using multinomial logistic regression. *BMC Infectious Diseases*, 20(1):485.

Yelin, I., Katz, R., Herzel, E., Berman-Zilberstein, T., Ben-Tov, A., Kuint, J., Gazit, S., Patalon, T., Chodick, G., and Kishony, R. (2021). Associations of the bnt162b2 covid-19 vaccine effectiveness with patient age and comorbidities. *medrxiv*, pages 2021–03.

Yoo, S. and Kim, J. (2016). Predicting college enrollment decision using multinomial logistic regression. *Journal of Applied Statistics*, 43(6):1093–1105.

Zhang, L., Peng, Y., Li, Y., Li, Y., Li, L., and Liang, H. (2014). Multinomial logistic regression analysis of factors affecting acceptance of pandemic a/h1n1 influenza vaccination among medical students in china. *PLoS ONE*, 9(5):e96552.