Constrained Statistical Inference for Categorical Data

by

Fares Said

A Thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy

Ottawa-Carleton Institute for
Mathematics and Statistics
(OCIMS)

Department of Mathematics and Statistics
Carleton University
Ottawa, Ontario, Canada
Wednesday 12th February, 2020

Copyright ©

2020 - Fares Said

The undersigned recommend to
the Faculty of Graduate Studies and Research
acceptance of the Thesis

Constrained Statistical Inference for Categorical Data

Submitted by **Fares Said**
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Dr. Sanjoy SINHA, Supervisor

Dr. Lang WU, External Examiner

Dr. Jose GALDO, Internal Examiner

Dr. Cai SONG, Institution Member

Dr. Chen XU, Institution Member

Dr. Mohamedou OULD HAYE, Defence Chair

Carleton University

2020

Abstract

Advancements in statistics are normally geared to addressing topics that will either address an existing gap in the field or to render analysis results more accurate/reliable. This work aims to add to existing research by extending from binary Generalized Linear Model (GLM) and Generalized Linear Mixed Model (GLMM) to a multinomial logit, multivariate GLM (MGLM) and multivariate GLMM (MGLMM), subject to ordered equality and inequality constraints. We extended the maximum likelihood estimate (MLE) and likelihood ratio hypothesis testing (LRT) methods for the binary and multinomial GLM and GLMM subject to linear equality and inequality constraints on the parameters of interest. These methods will build on existing literature to allow for more options in hypothesis testing and the construction of confidence intervals. The innovative procedures take advantage of the gradient projection (GP) technique for the MLE, and chi-bar-square statistics for constrained LRTs. The model presented in this thesis yields accurate results since parameter orderings or constraints often occur naturally; and when this occurs, we optimize the efficiency of a statistical method by incorporating the parameter constraints into the MLE and hypothesis testing. More specifically, we use ordered constrained inference for multinomial data whereby including equality and inequality constraints adds value to our predictions. Using real-world data from the Canadian Community Health Survey (C-CHS), the methodology of using constraints showed significant improvement on methodology that does not, which substantiates the added value of the work presented here.

This work contributes to the field by dealing with inequality constraints in MGLMM, specifically multinomial data, which is the most challenging problem in constrained inference. This helps improve results for researchers in both scientific and non-scientific fields.

Keywords: constrained/restricted statistical inference, optimization algorithms, gradient projection theory, quadratic programming, multinomial logit, projective geometry, convex cone.

Acknowledgments

As part of this thesis, I would like to take some time to thank all the people without whom this work would never have been possible. Although it is just my name on the cover, many people have contributed to the research in their own particular way, and for that I want to give them special thanks.

First and foremost, I would like to thank my thesis supervisors from the Department of Mathematics and Statistics at Carleton University: Dr. Chul Gyu Park, for his encouragement and support at the onset of this thesis; without his support, I would not have been able to begin this work. Dr. Sanjoy Sinha, without whom I would not have been able to stay focused; his advice and guidance helped shape this work into the final product you see here. The relationship we have cultivated over the past several years is one of genuine collaboration and respect, which I hope to continue even long after this thesis is presented. I must also take a moment to thank the Department of Mathematics and Statistics for all the opportunities it has afforded me over the years, including learning and teaching opportunities that helped me grow and gain the knowledge you see applied in this work, as well as the financial support through scholarships. Their assistance allowed me to continue my studies and research, and for this I am incredibly grateful.

Secondly, I would like to thank my government colleagues and peers at Immigration, Refugees and Citizenship Canada (IRCC) for their help in times of need. To Dr. Imran Ahmed, Dr. Somaieh Nikpoor, Elena Tipenko, and Abbas Rahal: thank you for your friendship, insightful comments and encouragement. To my peer and colleague at the Immigration Refugee Board (IRB): Alexandra Dykes: thank you for your continued words of encouragement, support and kindness over the past few years. And additional thanks goes to my colleagues at the Canada Border Services Agency (CBSA).

Thirdly, I would like to thank my family for their love, patience, and support while I completed this thesis, especially my mom, dad, and my wife. A very special thanks to my daughter, Anastasia, who gave me many moments of laughter and joy during the tough times. And to my son, Athanasius Atalla, who gave me the motivation to complete this thesis. Finally, I would like to thank God for His blessings and for the strength He has given me throughout this long and difficult process, without which none of this would be possible. I am thankful to God for all my accomplishments, especially this work.

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work.

This thesis has not been submitted for any other degree or for other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged as per acceptable referencing standards.

Fares Said

February 2020

Preface

This thesis is intended for statisticians, data scientists, applied researchers and students. It includes topics on categorical data analysis related to recent developments in the area of flexible and high-dimensional regression. This thesis develops a maximum likelihood inference subject to equality and inequality constraints for two important cases: regression parameter for MGLM and for MGLMM, with the logit link function, specifically for multinomial data.

We know that the unconstrained/unrestricted MLE has the following properties: it is consistent and asymptotically normally distributed, with the variance co-variance being the inverse of the Fisher information. However, for the constrained/restricted ML estimators, that property no longer holds. However, according to Hwang and Peddada [58], the distribution of the constrained estimator for linear models under simple ordering depends on how close the unconstrained estimator is to the boundary of the constraint.

Additional research into this aspect of MLEs could lead to discovering the asymptotic distribution for the restricted MLEs. This would allow further inference about these estimates and would be a good sequel to this research thesis. As regular advances to computational methods increase, studying Bayesian constrained techniques is a timely and useful research topic to pursue. Dunson and Neelon [59] highlight the importance of Bayesian constraints for GLMs. They note that sampling from the constrained posterior distribution is obtained by transforming draws from the unconstrained posterior density; this results in the direct application of the existing Gibbs sampling algorithms for posterior computation of GLMs.

Another consideration to expand on this work would be constraints on variance covariance parameters. Calvin and Dykstra [60] developed a residuals maximum likelihood estimation (REML) scheme for covariance matrices. Expanding on this research would help with GLMM where tests could be developed to identify trends in variance components.

Table of Contents

Abstract	iii
Acknowledgments	iv
Statement of Originality	vi
Preface	vii
Table of Contents	viii
List of Tables	xiii
List of Figures	xv
List of Acronyms	xvi
1 Introduction	1
1.1 Overview of Constrained Statistical Inference in GLM and GLMM	2
1.2 Statement of the Problem	3
1.3 Organisation of Thesis	4
2 Categorical Data Analysis	6
2.1 Introduction - Categorical Data	6
2.2 Distributions for Categorical Data	7
2.2.1 Poisson Experiment	7
2.2.2 Binomial Experiment	8
2.2.3 Multinomial Experiment	8
2.3 Estimation of Multinomial Probabilities	10

2.3.1	Distribution for MLE	11
2.4	Models for Two-dimensional Tables	13
2.4.1	Fixed Sample Size	15
2.4.2	Row-Fixed Sample Size	15
3	Regression Models for Categorical Response	17
3.1	Introduction	17
3.2	Logistic Regression for Binary Response	17
3.2.1	ML Estimation	19
3.2.2	Distribution for MLE	24
3.3	Logistic Regression for Multi-level Response	24
3.3.1	Nominal Responses: Baseline-Category Logit Models	26
3.3.2	Estimation of Model Parameters	29
3.4	Multinomial Logit Model as Multivariate GLM	36
3.4.1	Maximum Likelihood Estimation	37
3.4.2	Distribution for Multinomial Logit MLE	39
3.5	Multinomial Simulation Results	40
3.5.1	The Effect of Increase in Sample Size	44
4	Constrained Statistical Inference	46
4.1	Introduction	46
4.2	Concepts and Definitions	48
4.3	Constrained Optimization	51
4.3.1	Kuhn-Tucker(KT) Conditions	52
4.3.2	Gradient Projection Theory	55
5	Inference for Multivariate Normal under Linear Inequality Constraints	63
5.1	Order Restricted/Constrained Inference	63
5.2	Comparison of Population Order Means	67

5.2.1	Computing Restricted F and E Test	70
5.2.2	The Null Distribution of Restricted F-Test when $k=3$	72
5.2.3	The Null Distribution of Restricted F when k is more than 3	73
5.2.3.1	Computation of the exact p-value for the restricted F test	73
5.3	Constrained Tests on Multivariate Normal Mean	78
5.3.1	Likelihood Function	79
5.3.2	Constrained MLE and LRT	80
5.4	CHI-BAR-Square Distribution	89
5.4.1	CHI-BAR-SQUARE Weights	94
6	Inference for Categorical Data Under Linear Inequality Constraints	99
6.1	Generalized Linear Model	100
6.1.1	Unrestricted Inference in GLM	102
6.2	Restricted Estimation for Binary Data Using GP	103
6.2.1	Empirical Results for Constrained MLE for Binary Data	105
6.3	Constrained Tests for GLM	108
6.3.1	Empirical Results for Restricted LRT Under Binary GLM	113
6.4	GP Algorithm for Multinomial Logit Model	115
6.4.1	Simulation Study	115
6.5	Restricted MLE for Multinomial Logit Using GP	116
6.6	Restricted Tests for Multinomial Logit	122
7	Applications - Analysing CCHS Data Using Restricted Multinomial Logit	124
7.1	Canadian Community Health Survey	124
7.2	Description of the Asthma Subset of CCHS Data	125
7.3	Data Analysis	128
7.4	Restricted Inference	131

8	Constrained Statistical Inference in Multivariate GLMM for Multinomial	135
	Data	135
8.1	Introduction	135
8.2	Random Effects Models for Nominal Data	137
8.2.1	Model Specification	137
8.2.2	Baseline-Category Logit Models with Random Effects	138
8.3	Multivariate Likelihood Function	140
8.3.1	Maximum Likelihood	143
8.4	Random Intercept Multinomial Logit Model	145
8.4.1	Unconstrained ML Inference for CCHS Data	147
8.4.2	Previous Research	149
8.5	Constrained ML Inference for MGLMMs	149
8.5.1	Gradient Projection Algorithm for MGLMMs	149
8.5.2	Constrained Hypothesis Tests for MGLMMs	152
8.6	Constrained Statistical Inference for CCHS data	153
9	Conclusion	157
9.1	Main Findings	158
9.2	Limitations	159
9.3	Future Work	159
	Appendix	161
	List of References	161
	Appendix A Optimization Algorithms	168
A.1	Introduction	168
A.2	The Newton-Raphson Method	168
	Appendix B Exponential Family	170

Appendix C Linear Spaces	173
C.1 Vector Spaces	173
C.1.1 Subspaces, Linear Combinations, and Linear Varieties	174
C.1.2 Convexity and Cones	175
C.1.3 Linear Independence and Dimension	176
C.2 Normed Linear Space	177
C.2.1 Open and Closed Sets	178
C.2.2 Banach Spaces	179
C.3 Hilbert Space	180
Appendix D Big O and Small o	182
Appendix E Matrix Algebra	183
E.1 Matrix Differentiation	183
E.2 Projection	184
Appendix F R Code	185
Appendix G Distribution of Constrained MLEs for Multinomial Logit	187
G.1 Distribution for Case a, where all constraints are active	187
G.2 Distribution for Case b1, where at least one constraint is inactive	189
G.3 Distribution for Case b2, where at least one constraint is inactive	193
G.4 Distribution for Case c, where both constraints are inactive	197

List of Tables

2.1	Serum Cholesterol and Liver Disease	13
3.1	Voter counts for $N = 1200$	41
3.2	Voter counts for $N = 2400$	41
3.3	Bias, MSE, ECP and CIAW for $N = 1200$	42
3.4	Bias, MSE, ECP and CIAW for $N = 2400$	43
5.1	Size of Pituitary Fissure	64
5.2	Comparison of k means	68
5.3	Ordered Alternatives and ρ	72
5.4	The age at which a child first walks	75
5.5	The p -values for the $\bar{F}$ -test for different error distributions	78
6.1	Exponential Family of Distributions	101
6.2	Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for Bernoulli Model ($n = 100$)	106
6.3	Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for Bernoulli Model ($n = 300$)	107
6.4	Percentage of unrestricted MLE that satisfy the constraints	108
6.5	The empirical powers and sizes of restricted and unrestricted LRT for $n = 100$ and $n = 300$ at 5% significance level	114
6.6	Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for MN Logit Model ($N = 350$)	118
6.7	Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for MN Logit Model ($N = 700$)	119
6.8	Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for MN Logit Model ($N = 1000$)	120

6.9	Percentage of unrestricted MLE that satisfy the constraints	121
6.10	Empirical powers and sizes of restricted and unrestricted LRT for $N = (250, 350,$ 700, and 1000) at 5% significance level	123
7.1	Asthma Subset from CCHS Data	127
7.2	Summary Statistics of Asthma from CCHS	128
7.3	Unrestricted MLE for multinomial logit	129
7.4	Unrestricted and Restricted MLE for multinomial logit of Asthma	133
8.1	Unrestricted MLE for multinomial logit	148
8.2	Unrestricted and Restricted MLE for random intercept multinomial logit of Asthma	155
B.1	Exponential Distribution	172

List of Figures

3.1	S-shaped: Simple logistic probability distribution	19
3.2	Kernel density and histograms of unconstrained MLEs $\hat{\beta}_{ij}$	45
4.1	Polyhedron $\mathcal{P}$ (shown shaded) is the intersection of five half-spaces, with outward normal vectors $a_1, \dots, a_5$	49
5.1	Geometry of constrained LRT	65
5.2	Two dimensions constrained MLE of θ subject to $\mathbf{A}\theta \geq \mathbf{0}$, and the LRT of H_0 vs H_1	81
5.3	Two dimensions constrained MLE of θ subject to $\theta \geq \mathbf{0}$, and the LRT of H_0 vs H_1	84
5.4	The constrained MLE of θ subject to $\theta \in \mathcal{C}$ and the LRT of H_0 vs H_1 and a typical boundary of the critical region is PQRS	88
5.5	OB and OC are the $\mathbf{V}$ -projections of OA onto $\mathcal{C}$ and $\mathcal{C}^\circ$ respectively.	89
C.1	Cone	175
G.1	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case a	188
G.2	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b ₁	190
G.3	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b ₁	191
G.4	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b ₁	192
G.5	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b ₂	194
G.6	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b ₂	195
G.7	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b ₂	196
G.8	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case c	198
G.9	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case c	199
G.10	Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case c	200

List of Acronyms

Acronyms	Definition
AGQ	Adaptive Gauss-Hermite Quadrature
CCHS	Canadian Community Health Survey
cdf	Cumulative Density Function
CIAW	Confidence Interval Average Width
CIHI	Canadian Institute for Health Information
CL	Conditional Logit
CLT	Central Limit Theorem
CP	Conservative Party
CSI	Constrained Statistical Inference
ECP	Estimated Coverage Probability
EF	Exponential Family
EFS	Empirical Fisher Scoring
EM	Expectation Maximization
FS	Fisher Scoring
GEE	Generalized Estimating Equations
GH	Gauss-Hermite

GL	Generalized Logit
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GP	Gradient Projection
iid	Identically Independently Distributed
IRWLS-QP	Iteratively Reweighted-Least Squares-Quadratic Programming
KKT	Karush-Kuhn-Tucker
KT	Kuhn-Tucker
LP	Liberal Party
LRT	Likelihood Ratio Test or Likelihood Ratio Hypothesis Test
MGLM	Multivariate Generalized Linear Models
MGLMM	Multivariate Generalized Linear Mixed Models
ML	Mixed Logit
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator/Estimation/Estimate
MNLR	Multinomial Logistic Regression
MSE	Mean Squared Error
NR	Newton-Raphson
pdf	probability density function

REML	Residuals Maximum Likelihood Estimation
RMLE	Restricted Maximum Likelihood Estimation
RSS	Residual Sum Square
SRS	Simple Random Sample
w.r.t.	with respect to

Chapter 1

Introduction

Over the past decade, my educational experiences led to an internship at the Canadian Institute for Health Information (CIHI), a career with the federal public service as a statistician/data scientist in various departments such as Health Canada, Immigration, Refugees, and Citizenship Canada (IRCC), the Immigration and Refugee Board of Canada (IRB), the Canada Border Services Agency (CBSA), and a long-standing relationship with the Department of Mathematics and Statistics at Carleton University, both as a teaching assistant and a course instructor. I have experienced first-hand in a wide-ranging set of environments the lack of and strong need for literature and research into modelling using constrained inference with multinomial data. This thesis develops a maximum likelihood inference subject to equality and inequality constraints for two important cases: regression parameter for the multivariate generalized linear model (MGLM) and for the multivariate generalized linear mixed model (MGLMM), with the logit link function, specifically for multinomial data. This differs from existing works in that it considers multinomial data (extension from the logit model to multivariate logit settings), not just binary data. For this reason, the gradient projection algorithm is implemented for obtaining maximum likelihood estimators of regression parameters for multinomial data. These estimators are then used in constrained likelihood ratio tests. The asymptotic null distribution of the constrained likelihood ratio tests is also derived and is found to be a chi-bar-square (a mixture of chi-square distributions). Empirical results are obtained using simulations to compare methods of estimation and testing. Finally, real-world applications are considered as part of the Canadian Community Health Survey (CCHS) analysis.

This thesis includes topics on categorical data analysis related to recent developments in the area of flexible and high-dimensional regression. Readers of this thesis should have background

in regression, maximum likelihood methods, mathematical and statistical theories, as well as interest in constrained inference. Those with minimal background in theories should still be comfortable in following the methodologies applied to the CCHS in Chapter (7).

1.1 Overview of Constrained Statistical Inference in GLM and GLMM

In many statistical applications, we use the GLM when the mean of an observation is not a linear combination of parameters; when it is linked with a linear function of parameters of interest to various explanatory variables through some nonlinear function, called the link function; and when the data are not normally distributed, but follow a distribution in the exponential family (EF); and when the variance of the response is not constant but a function of the mean. To address overdispersion and correlation in the model, we can incorporate random effects, and in doing so, we rename the model to the GLMM. For additional details on GLM, see Section 6.1. Generalized linear mixed models are used to model clustered and longitudinal data in which the distribution of the response variable is a member of the EF (see Appendix B for information on EF). These models have found applications in various research and development fields (epidemiology, genetics, biology, market research, economics, security, etc.). Examples include biostatistics, health sciences, medical treatments, econometrics, fraud detection, etc.

GLMMs consist of both fixed and random effects parameters. Fixed effects parameters relate covariates to the response at the population level. Random effects parameters relate covariates to the response at the individual level. GLMM identifies clusters of data based on similarities among the random effects parameters. For additional details on GLMM, refer to Chapter 8.

For GLM and GLMM, using constraints has numerous benefits (covered in detail in Chapter 4, Section 4.1). Using constraints requires more complex algorithms and computational power, meaning more time to implement computations and more case-specific algorithms. However, they may be deemed more useful given the improved efficiency and accuracy of the results.

1.2 Statement of the Problem

As with all forms of advancement, constrained statistical inference has followed a progressive path over the decades, starting in its early years with the pioneer work of Constance Van Eeden [63], whose work contributed to maximum likelihood estimation techniques and the collaborative work of D.J. Bartholomew, R.E. Barlow, H.D. Brunk and J.M. Bremner on isotonic regression [64]. This was closely followed by work on inferences under normal or multinomial settings by a number of scholars, namely Akio Judo (1963) and Mervyn Silvapulle (1994) (with contributions to the one-sided test) [68] and [69], Richard L. Dykstra (with his development of an algorithm for restricted least squares regression), Hammou El Barmi in collaboration with Dykstra (who proposed a method for fitting models involving both convex and log-convex constraints on the probability vectors of a product multinomial distribution) [65], and many others. More recently, works by Lin (with contribution on variance component testing in GLMs with random effects) [66], built upon by work from Hall and Praestgaard (with contribution on order restricted score tests for homogeneity in generalised linear and nonlinear mixed models) [67]. Many more scholars have since contributed various aspects to this field, and have demonstrated through simulations and the testing of theories that the proper consideration for constraints in modelling allows for increased testing power and better accuracy in predictions.

The progression matched the needs of the day, with modelling for unconstrained binary data being sufficient at the onset and then progressively requiring constrained binary data modelling. As time elapsed, technological developments generated a need for more complex modelling techniques to be explored such as multinomial modelling. Due to industry needs for predictions of multi-categorical responses, and with the popularisation of artificial intelligence, machine learning, and data science (resulting in computational sufficiency), additional work in constrained statistical inference became both relevant and important. To date, little research has been conducted for the advancement and development of multi-level categorical responses. This thesis addresses this problem while also expanding its usefulness with the addition of constraints in multinomial logit models for MGLM and MGLMM.

1.3 Organisation of Thesis

The ideas, theories and methods described in this thesis are intended for those with backgrounds in the field. The information is presented using a simple to complex method, meaning we progress from simple ideas to more complex ones, i.e. from binary GLM to multinomial, multivariate GLM and multivariate GLMM. An overview of categorical data concepts is provided as a starting point in Chapter 2. More advanced concepts related to modeling techniques are presented in Chapter 3, where the Newton-Raphson technique is implemented to find unrestricted MLE for the multinomial logit model, where simulations were conducted to study empirical properties of the estimators. Chapter 3, Section 3.5 presents the results of these simulations, which verifies the validity and performance of the algorithm by showing consistent results with the MLE properties. An overview of constrained statistical inference is presented in Chapter 4, which also covers definitions of convex set and convex cone, and discusses how to derive Kuhn-Tucker (KT) conditions and how the Gradient Projection algorithm is modified to satisfy the needs of this thesis in handling clustered correlated multinomial data.

Chapters 2 through 4 prepare the reader with all the background information needed to understand the concepts presented in Chapter 5 and 6. These two chapters, combined with Chapter 8, comprise the bulk of this thesis. Chapter 5 covers constrained inference under normal data that uses the $\bar{F}$ -test for the mean comparison and derives the chi-bar-square distribution. Chapter 6 uses the NR technique and the modified Gradient Projection (GP) algorithm to find the restricted MLEs for GLM binary data and the restricted MLEs for MGLM multinomial data. Also, we derive the asymptotic distribution for the restricted likelihood ratio test, which follows a chi-bar-square distribution. After conducting simulations for the GLM and MGLM, we found that restricted MLEs have larger bias and smaller mean squared error (MSE) than the unrestricted counterparts and that the bias and the MSE decreases as the sample sizes increase. If data are obtained from parameters within the cone formed by the constraints, not around the boundary, then the restricted and unrestricted MLEs are often the same (between 70% to 90% of the time). We also find that the restricted likelihood ratio tests provide acceptable

empirical size and better power performance than the unrestricted likelihood ratio test when the constraints are satisfied (refer to Tables (6.5) and (6.10) for details). Chapter 7 implements the theory developed and tested in Chapters 5 and 6 and applies it to real-world data from the Canadian Community Health Survey (CCHS). Finally, Chapter 8 covers the constrained statistical inference for multivariate GLMM where ordered equality and inequality constraints are imposed on the multinomial logit model. Particular attention is given to ordered inequality constraints in multivariate GLMM as this is the most challenging problem in constrained inference, whereas estimation with equality constraints is fairly straightforward. Here we also extend the multivariate GLM with the multinomial logit to the multivariate GLMM. The method is applied to the CCHS data by treating the effects of regions as intercept random effects with one variance component.

As an added support to this thesis, readers can benefit from referencing the appendices which provide details on various topics, including optimization algorithms, exponential family (EF), basic properties of vector space and normed linear spaces, matrices and vectors, the R code developed for this thesis (which has been added to GitHub; see Appendix (F) on page 185), and the detailed results for the numerical studies presented throughout the thesis.

Chapter 2

Categorical Data Analysis

2.1 Introduction - Categorical Data

A *categorical variable* has a measurement scale consisting of a set of categories. For instance, political philosophy is often measured as liberal, moderate, or conservative. Diagnoses regarding breast cancer based on a mammogram use the categories normal, benign, probably benign, suspicious, and malignant. Substances are defined as solid, liquid or gas; and so on. Categorical variables have two primary types of scales: nominal or ordinal.

- (1) Variables with categories that do not follow a natural order are called **nominal**. For nominal variables, the order of listing the categories is irrelevant, and the statistical analysis does not depend on that ordering. Examples are:

- * religious affiliations (Catholic, Protestant, Jewish, Muslim, other),
- * mode of transportation to work (automobile, bicycle, bus, subway, walk),
- * favorite type of music (classical, country, folk, jazz, rock), and
- * choice of residence (apartment, condominium, house, other).

- (2) Many categorical variables have ordered categories. These variables are called **ordinal**. Ordinal variables have ordered categories, but distances between the categories are unknown [1]. Examples are:

- * size of automobile (subcompact, compact, midsize, large),
- * social class (upper, middle, lower),
- * political philosophy (liberal, moderate, conservative), and

* patient condition (good, fair, serious, critical).

Nominal variables are qualitative, where distinct categories differ in quality, not in quantity. Interval variables are quantitative, where distinct levels have differing amounts of the characteristic of interest [1].

2.2 Distributions for Categorical Data

Inferential data analysis requires assumptions about the random mechanism that generated the data. For continuous responses, the normal distribution plays the central role. In this section, we review the three key distributions for discrete and categorical responses (i.e. the outcome of each experiment belongs to exactly one of c categories):

- (1) Poisson (2) If $c = 2$, Binomial (3) If $c \geq 3$, Multinomial

2.2.1 Poisson Experiment

Sometimes count data do not result from a fixed number of trials. For instance, if Y = number of deaths due to automobile accidents on motorways in Italy during this coming week, there is no fixed upper limit n for Y (as you are aware if you have driven in Italy). Since Y must be a nonnegative integer, its distribution should place its mass on that range. The simplest such distribution is the Poisson. Its probabilities depend on a single parameter, the mean μ , where

$$P(Y = k) = \frac{e^{-\mu} \mu^k}{k!}, \quad k = 0, 1, \dots$$

The Poisson distribution is used for counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent. It also applies to an approximation for the binomial when n is large and p is small, with $\mu = np$ [1].

2.2.2 Binomial Experiment

A binomial experiment is one that has the following properties:

- (1) The experiment consists of n identical trials.
- (2) Each trial results in one of two outcomes. We will label one outcome a success and the other a failure.
- (3) The probability of success on a single trial is π , which remains the same from trial to trial.
- (4) The trials are independent, i.e. the outcome of one trial does not influence the outcome of any other trial.
- (5) The random variable Y is the number of successes observed in n trials.

The probability of observing Y successes in n trials of a binomial experiment is

$$P(Y = k) = \frac{n!}{k!(n-k)!} \pi^k (1-\pi)^{n-k},$$

for $k = 0, 1, \dots, n$. The binomial distribution for Y possesses a mound-shaped probability distribution that can be approximated by using a normal curve when

$$n \geq \frac{5}{\min(\pi, 1-\pi)} \quad (\text{or equivalently,}) \quad n\pi \geq 5 \text{ and } n(1-\pi) \geq 5.$$

Note 2.1 (Inferences): *Using a binomial experiment, we can conduct inferences about one population proportion π or the difference between two population proportions $\pi_1 - \pi_2$ [1].*

2.2.3 Multinomial Experiment

We can conduct trials where the result is more than two possible outcomes. In these cases, suppose that each of n identical and independent trials can have an outcome in any of c categories. Here we can extend the binomial sampling scheme of Section 2.2.2 to situations in

which each trial results in one of c possible outcomes, where ($c > 2$) [1]. This type of experiment is called a multinomial experiment with the following characteristics:

- (1) Let $y_{ij} = \begin{cases} 1 & \text{if trial } i \text{ has an outcome in category } j \\ 0 & \text{otherwise} \end{cases}$, with $\sum_{j=1}^c y_{ij} = 1$ and $n_j = \sum_{i=1}^n y_{ij}$.
- (2) The experiment consists of n identical and independent trials.
- (3) Each trial results in one of c outcomes.
- (4) The probability that a single trial has an outcome in category j is $\pi_j = P(Y_{ij} = 1)$ for $j = 1, 2, \dots, c$, and π_j remains **constant** from trial to trial. (Note: The sum of all c probabilities $\sum_{j=1}^c \pi_j = 1$).
- (5) We are interested in the number of outcomes n_j in each category j . (Note: $\sum_{j=1}^c n_j = n$).

We obtain the multinomial distribution by drawing a simple random sample (SRS) of size n from the population with c categories. We then classify our categories and summarize our sample using the following table:

	Categories						
	1	2	$\dots$	j	$\dots$	c	Totals
Cell Probabilities	π_1	π_2	$\dots$	π_j	$\dots$	π_c	1
Obs. Frequencies	n_1	n_2	$\dots$	n_j	$\dots$	n_c	n
$\mathbf{y}_1$	y_{11}	y_{12}	$\dots$	y_{1j}	$\dots$	y_{1c}	1
$\mathbf{y}_2$	y_{21}	y_{22}	$\dots$	y_{2j}	$\dots$	y_{2c}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}_i$	y_{i1}	y_{i2}	$\dots$	y_{ij}	$\dots$	y_{ic}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}_n$	y_{n1}	y_{n2}	$\dots$	y_{nj}	$\dots$	y_{nc}	1

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ic})$ represents a multinomial trial. The counts $(n_1, n_2, \dots, n_c)$ have a *multinomial distribution* with the following probability mass function:

$$P(n_1, n_2, \dots, n_c) = \frac{n!}{n_1! n_2! \dots n_c!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_c^{n_c} \quad (2.1)$$

subject to the constraints

$$\sum_{j=1}^c n_j = n \text{ and } \sum_{j=1}^c \pi_j = 1,$$

where

* n_j and π_j are the number of outcomes and probability of success on a single trial in category j , respectively.

* $E(n_j) = n\pi_j$, $V(n_j) = n\pi_j(1 - \pi_j)$ and $\text{Cov}(n_j, n_\ell) = -n\pi_j\pi_\ell$ with $\ell \neq j = 1, \dots, c$.

2.3 Estimation of Multinomial Probabilities

The joint probability of a vector $(n_1, n_2, \dots, n_c)$ is called the multinomial, and has the form:

$$P(n_1, n_2, \dots, n_c) = f(n_1, n_2, \dots, n_c | \pi_1, \pi_2, \dots, \pi_c) = \frac{n!}{\prod_{j=1}^c n_j!} \prod_{j=1}^c \pi_j^{n_j}.$$

We can maximize the likelihood function to obtain estimators of the parameters π_j . The *log-likelihood* is given by

$$\ell(\boldsymbol{\pi}) = \ell(\pi_1, \pi_2, \dots, \pi_c) = \ln(n!) - \sum_{j=1}^c \ln(n_j!) + \sum_{j=1}^c n_j \ln(\pi_j).$$

However, we cannot just move forward and maximize this. To maximize $\ell(\boldsymbol{\pi})$ subject to the constraint $\sum_{j=1}^c \pi_j = 1$, we can use Lagrange's multiplier:

$$L(\pi_1, \pi_2, \dots, \pi_c, \lambda) = \ell(\pi_1, \pi_2, \dots, \pi_c) - \lambda \left(\sum_{j=1}^c \pi_j - 1 \right),$$

where λ is called Lagrange's Multiplier. To maximize $L(\cdot)$, we take the partial derivatives and set them equal to zero. We have

$$\frac{\partial L}{\partial \pi_j} = \frac{n_j}{\pi_j} - \lambda \text{ and } \frac{\partial L}{\partial \lambda} = - \left(\sum_{j=1}^c \pi_j - 1 \right) = 1 - \sum_{j=1}^c \pi_j.$$

Setting $\frac{\partial L}{\partial \pi_j} = \frac{\partial L}{\partial \lambda} = 0$ yields

$$\frac{n_j}{\hat{\pi}_j} - \hat{\lambda} = 0 \Rightarrow \hat{\pi}_j = \frac{n_j}{\hat{\lambda}} \Rightarrow n_j = \hat{\pi}_j \hat{\lambda} \text{ and } \sum_{j=1}^c \hat{\pi}_j - 1 = 0 \Rightarrow \sum_{j=1}^c \hat{\pi}_j = 1.$$

Since $n = \sum_{j=1}^c n_j$ and $n_j = \hat{\pi}_j \hat{\lambda}$, thus $n = \sum_{j=1}^c n_j = \hat{\lambda} \sum_{j=1}^c \hat{\pi}_j \Rightarrow n = \hat{\lambda}$; therefore the MLE for π_j is

$$\hat{\pi}_j = \frac{n_j}{\hat{\lambda}} = \frac{n_j}{n} \quad (2.2)$$

since the second derivative $\frac{\partial^2 L}{\partial \pi_j^2} = -\frac{n_j}{\pi_j^2} = -\frac{n^2}{n_j}$ is negative.

2.3.1 Distribution for MLE

Consider the population probability column vector:

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_c)^T,$$

and MLE probability column vector:

$$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_c)^T.$$

Consider the i^{th} trial outcome:

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{ic})^T,$$

where Y_{ij} defined above is:

$$Y_{ij} = \begin{cases} 1 & \text{if trial } i \text{ has outcome in category } j \\ 0 & \text{otherwise.} \end{cases}$$

Since each observation falls in just one cell, $\sum_{j=1}^c Y_{ij} = 1$ for each $i = 1, \dots, n$ and $Y_{ij}Y_{i\ell} = 0$ when $j \neq \ell$. Also, $\hat{\pi}_j = \frac{n_j}{n} = \frac{\sum_{i=1}^n Y_{ij}}{n}$, and the characteristics for Y_{ij} are

$$E(Y_{ij}) = 0 \times P(Y_{ij} = 0) + 1 \times P(Y_{ij} = 1) = \pi_j = E(Y_{ij}^2) \text{ and } E(Y_{ij}Y_{i\ell}) = 0.$$

Thus,

$$\sigma_{jj} = V(Y_{ij}) = E(Y_{ij}^2) - [E(Y_{ij})]^2 = \pi_j - \pi_j^2 = \pi_j(1 - \pi_j)$$

$$\sigma_{j\ell} = \text{Cov}(Y_{ij}, Y_{i\ell}) = E(Y_{ij}Y_{i\ell}) - E(Y_{ij})E(Y_{i\ell}) = 0 - \pi_j\pi_\ell = -\pi_j\pi_\ell \text{ for } j \neq \ell.$$

Using these results, we then can write the mean vector and covariance matrix of $\mathbf{Y}_i$ as

$$E(\mathbf{Y}_i) = \boldsymbol{\pi} \text{ and } \text{Cov}(\mathbf{Y}_i) = E(\mathbf{Y}_i - \boldsymbol{\pi})(\mathbf{Y}_i - \boldsymbol{\pi})^T = E(\mathbf{Y}_i\mathbf{Y}_i^T) - E(\mathbf{Y}_i)E(\mathbf{Y}_i^T) = \Sigma,$$

where

$$\begin{aligned} \Sigma &= \begin{pmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_c \\ -\pi_2\pi_1 & \pi_2(1 - \pi_2) & \cdots & -\pi_2\pi_c \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_c\pi_1 & -\pi_c\pi_2 & \cdots & \pi_c(1 - \pi_c) \end{pmatrix} = \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_c \end{pmatrix} - \begin{pmatrix} \pi_1\pi_1 & \pi_1\pi_2 & \cdots & \pi_1\pi_c \\ \pi_2\pi_1 & \pi_2\pi_2 & \cdots & \pi_2\pi_c \\ \vdots & \vdots & \ddots & \vdots \\ \pi_c\pi_1 & \pi_c\pi_2 & \cdots & \pi_c\pi_c \end{pmatrix} \\ &= \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T. \end{aligned}$$

Since $\hat{\pi}_j = \frac{n_j}{n} = \frac{\sum_{i=1}^n Y_{ij}}{n}$, and the vector $\hat{\boldsymbol{\pi}} = \frac{\sum_{i=1}^n \mathbf{Y}_i}{n}$ is a sample mean of n independent observations, the mean vector and covariance matrix of $\hat{\boldsymbol{\pi}}$ are

$$\mathbf{E}(\hat{\boldsymbol{\pi}}) = \mathbf{E}\left(\frac{\sum_{i=1}^n \mathbf{Y}_i}{n}\right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(\mathbf{Y}_i) = \boldsymbol{\pi},$$

and

$$\text{Cov}(\hat{\boldsymbol{\pi}}) = \mathbf{E}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})^T = \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \mathbf{E}(\mathbf{Y}_i - \boldsymbol{\pi})(\mathbf{Y}_k - \boldsymbol{\pi})^T = \frac{1}{n^2} n \Sigma = \frac{\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T}{n}.$$

Note 2.2: This covariance matrix is **singular**, because of the linear dependence $\sum_{j=1}^c \pi_j = 1$.

Using the multivariate Central Limit Theorem, we can write that

$$\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{d} \mathbf{Z} \sim N_c(\mathbf{0}, \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T) \quad (2.3)$$

By the *delta method*, functions of $\hat{\boldsymbol{\pi}}$ having nonzero differential at $\boldsymbol{\pi}$ are also asymptotically normal.

2.4 Models for Two-dimensional Tables

We start by considering the simplest possible contingency table: 2×2 table. Suppose that Table 2.1 is based on a longitudinal study of liver disease. It shows 1430 patients cross-classified by the level of their serum cholesterol (below or above 260) and the presence or absence of liver disease.

Table 2.1: Serum Cholesterol and Liver Disease

Serum Cholesterol	Liver Disease		Total
	Present	Absent	
<260	63	1005	1068
260+	49	313	362
Total	112	1318	1430

Speaking in a more general context, we let X and Y denote two categorical variables, where X is a row factor with r categories indexed by i and Y is a column factor with c categories indexed by j . This forms an $r \times c$ contingency table where the classifications of subjects on X and Y have rc possible combinations.

- * If both X and Y are response variables, we study their *joint distribution*, and can also compute their marginal and conditional distributions.
- * If Y is a response variable and X is an explanatory variable, we study the *conditional distribution* of Y and how it changes as the category of X changes.
- * The *cells* of the table represent rc possible outcomes, which are frequency counts of outcomes from a random sample of subjects taken from a particular population.
- * Let $\pi_{ij} = P(X = i, Y = j)$ denote the probability that (X, Y) occurs in the cell of row i and column j .
- * Let $\pi_{i.} = P(X = i) = \sum_{j=1}^c \pi_{ij}$ denote the *marginal* probability that the row variable (X) takes the value i , and let $\pi_{.j} = P(Y = j) = \sum_{i=1}^r \pi_{ij}$ denote the *marginal* probability that the column variable (Y) takes the value j , with constraints:

$$\sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = \sum_{j=1}^c \pi_{.j} = \sum_{i=1}^r \pi_{i.} = 1.$$

- * The cell frequencies are denoted by n_{ij} , $n = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$ is the total sample size, $n_{i.} = \sum_{j=1}^c n_{ij}$ row total, and $n_{.j} = \sum_{i=1}^r n_{ij}$ column total.

When X is fixed and Y is a random response variable, the assumption of a joint distribution for X and Y no longer applies. Instead, we would look at how Y changes as the category of X changes. Given that a subject is classified in row i of X , we use $\pi_{j|i} = P(Y = j | X = i)$ to denote the conditional probability of classification in column j of Y at various levels of explanatory variables.

2.4.1 Fixed Sample Size

Let Y_{ij} denote a random variable that represents the number of observations in (i, j) -th cell, with an observed value y_{ij} . When the total sample size n is fixed, but the row and column totals are not, a multinomial sampling model applies. The joint distribution of the counts is then the multinomial distribution, with the probability mass function (pmf):

$$P(\mathbf{Y} = \mathbf{y}) = \frac{n!}{y_{11}! \cdots y_{rc}!} \pi_{11}^{y_{11}} \cdots \pi_{rc}^{y_{rc}} = \frac{n!}{y_{11}! \cdots y_{rc}!} \prod_{i=1}^r \prod_{j=1}^c \pi_{ij}^{y_{ij}}, \quad (2.4)$$

where $\mathbf{Y}$ is a random vector collecting all rc counts, and $\mathbf{y}$ is a vector of observed values. We obtain the kernel multinomial log-likelihood function by taking the natural logs, which for a general $r \times c$ table has the form

$$\ln L = \sum_{i=1}^r \sum_{j=1}^c y_{ij} \ln(\pi_{ij}),$$

subject to the constraints:

$$\sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = \sum_{j=1}^c \pi_{.j} = \sum_{i=1}^r \pi_{i.} = 1.$$

This restriction may be imposed by adding a Lagrange multiplier, or by writing the last probability as the complement of all others. Then we can estimate the parameters by taking the derivatives of the log-likelihood function with respect to π_{ij} . The unrestricted maximum likelihood estimators are obtained as:

$$\hat{\pi}_{ij} = \frac{y_{ij}}{n}.$$

2.4.2 Row-Fixed Sample Size

Consider a random variable Y_i that may fall in category j with the probability

$$\pi_{ij} = P(Y_i = j). \quad (2.5)$$

- * When observations on a response Y occur separately at each setting i of an explanatory variable X , we treat row totals as fixed. For simplicity, we use the notation $n_i = n_{i\cdot}$, and suppose that n_i observations on Y at setting i of X are independent, each with probability distribution $\{\pi_{i1}, \dots, \pi_{ic}\}$.
- * Assuming that the response categories are mutually exclusive, we have $\sum_{j=1}^c \pi_{ij} = 1$ for each i and we have only $c - 1$ parameters.

Let n_i denote the number of cases in the i -th subject/level and let Y_{ij} denote the number of responses from the i -th subject that fall in the j -th category, with observed value $y_{ij} = n_{ij}$. The probability distribution of counts Y_{ij} satisfying $\sum_{j=1}^c n_{ij} = n_i$ then have multinomial distribution:

$$P(\mathbf{Y}_i = \mathbf{y}_i) = P(Y_{i1} = y_{i1}, \dots, Y_{ic} = y_{ic}) = \frac{n_i!}{\prod_{j=1}^c y_{ij}!} \prod_{j=1}^c \pi_{ij}^{y_{ij}}.$$

The joint probability function for the entire data set is the product r levels of the multinomial function from various settings such that:

$$f(\mathbf{y}) = \prod_{i=1}^r \left[\frac{n_i!}{\prod_{j=1}^c y_{ij}!} \prod_{j=1}^c \pi_{ij}^{y_{ij}} \right] \quad (2.6)$$

since samples at different settings i of X are independent.

The expected value of $E(\mathbf{Y}_i) = n_i \boldsymbol{\pi}_i$ and the covariance $V(\mathbf{Y}_i) = n_i(\text{Diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^T)$, where

$$\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{ic})^T.$$

Chapter 3

Regression Models for Categorical Response

3.1 Introduction

In Chapter 2, we focused on methods to estimate and make inference about the probability of success π_{ij} using the contingency tables. Most studies, however, model these probabilities as the function of a vector $\mathbf{x}_i$ of the covariates associated with the i -th individual, subject, or group. The logistic model is the most popular regression model to characterize the relationship between a categorical dependent variable and a set of independent variables (or predictors, covariates, etc.). The dependent variable in logistic regression is binary (or dichotomous), but it can be a multi-level polytomous outcome with more than two response levels in the general case. Various sections in this chapter are inspired by a variety of sources including works by Alan Agresti, Scott A. Czepiel, and others [23], [2], [25], and [26].

3.2 Logistic Regression for Binary Response

Consider a sample of n subjects. For each subject, let

- (1) Y_i denote a binary response of interest for the i -th subject taking two values $(0, 1)$,
- (2) $\mathbf{x}_i = (x_{i0}, x_{i1}, \dots, x_{ip})^T$ denote a column vector of independent variables for the i -th subject.

Assume

$$Y_i|\mathbf{x}_i \sim \text{Bernoulli}(\pi_i); \quad E(Y_i|\mathbf{x}_i) = \pi_i = \pi(\mathbf{x}_i) = P(Y_i = 1|\mathbf{x}_i).$$

Thus, $P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$ for $y_i = 0, 1$ is the PMF for Y_i .

The likelihood function is

$$L(\boldsymbol{\pi}|y_1, \dots, y_n) = \prod_{i=1}^n P(Y_i = y_i) = \prod_{i=1}^n \pi_i^{y_i}(1 - \pi_i)^{1-y_i}.$$

The simplest type of function for $\pi(\mathbf{x}_i)$ is the linear model:

$$\pi(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} = \mathbf{x}_i^T \boldsymbol{\beta}.$$

However, this model could lead to values of π_i less than 0 or greater than 1, depending on the values of the explanatory variables and regression parameters. Fortunately, many non-linear expressions are available that force π_i to be between 0 and 1. The most commonly used expression is the logistic regression model:

$$\pi(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \text{ which leads to } 0 \leq \pi_i \leq 1.$$

The logistic regression has the following general form:

$$\text{logit}(\pi_i) = \log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (3.1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ is the vector of parameters for the independent variables with $x_{i0} = 1$ associated to the intercept β_0 . In the logistic model, we are modeling the effect of $\mathbf{x}$ on the response rate by relating $\text{logit}(\pi_i)$ or log odds of response $\log \left(\frac{\pi_i}{1 - \pi_i} \right)$ to a linear function of $\mathbf{x}$ of the form:

$$\eta_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Consider the simple logistic probability distribution given by:

$$f(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}, \quad -\infty < x < \infty,$$

whose plot can be modeled by an S-shaped (forward or backward, depending on the sign of x coefficient) sigmoidal curve given by:

Figure 3.1: S-shaped: Simple logistic probability distribution

3.2.1 ML Estimation

The likelihood function for equation (3.1) is:

$$\begin{aligned} L(\boldsymbol{\beta}|\mathbf{y}) &= \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \prod_{i=1}^n \left[\left(\frac{\pi_i}{1-\pi_i} \right)^{y_i} (1 - \pi_i) \right] \\ &= \prod_{i=1}^n \left[\left(e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right)^{y_i} \left(1 - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) \right] = \exp \left(\sum_{i=1}^n y_i \mathbf{x}_i^T \boldsymbol{\beta} \right) \prod_{i=1}^n \frac{1}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}. \end{aligned}$$

The log-likelihood function is:

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \log \left\{ \exp \left(\sum_{i=1}^n y_i \mathbf{x}_i^T \boldsymbol{\beta} \right) \prod_{i=1}^n \frac{1}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right\} = \sum_{i=1}^n y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^n \log(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}).$$

We take the derivatives w.r.t $\beta_0, \dots, \beta_p$, set these equal to 0, and solve them simultaneously to obtain the parameter estimates $\hat{\beta}_0, \dots, \hat{\beta}_p$. Unfortunately, there are only a few simple cases where these parameter estimates have closed-form solutions. Instead, we use iterative numerical procedures computed by the Newton-Raphson (NR) method which requires finding a stationary point of the gradient of the log-likelihood to solve the optimization problem. To maximize $\ell(\boldsymbol{\beta})$, we compute the score/gradient function which is given by:

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

so we are solving a system of $p + 1$ non-linear equations. Let us now compute $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}$ where β_j is a j -th element of $\boldsymbol{\beta}$. It is important to realize that $\mathbf{x}_i$ presents a linear relationship between $\ell(\boldsymbol{\beta})$ the elements of $\boldsymbol{\beta}$. Thus each of the partial derivatives in $S(\boldsymbol{\beta})$ will have the same form:

$$S(\beta_j) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^n \left\{ y_i \frac{\partial (\mathbf{x}_i^T \boldsymbol{\beta})}{\partial \beta_j} - \frac{\partial \log(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}})}{\partial \beta_j} \right\},$$

where

$$\frac{\partial (\mathbf{x}_i^T \boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial (\beta_0 + \beta_1 x_{i1} + \dots + \beta_j x_{ij} + \dots + \beta_p x_{ip})}{\partial \beta_j} = x_{ij}, \quad \text{where } x_{i0} = 1$$

and

$$\frac{\partial \log(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}})}{\partial \beta_j} = \frac{\frac{\partial \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\partial \beta_j}}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \frac{\partial (\mathbf{x}_i^T \boldsymbol{\beta})}{\partial \beta_j} = \pi(\mathbf{x}_i) x_{ij} = \pi_i x_{ij}.$$

So,

$$S(\beta_j) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^n \{y_i x_{ij} - \pi_i x_{ij}\} = \sum_{i=1}^n \{x_{ij}(y_i - \pi_i)\}, \quad j = 0, \dots, p.$$

The vector form for score equation is:

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n (y_i - \pi_i) \mathbf{x}_i.$$

Since $\frac{\partial \pi_i}{\partial \boldsymbol{\beta}} = \pi_i(1 - \pi_i)\mathbf{x}_i$, the partial derivatives for the k -th element is $\frac{\partial \pi_i}{\partial \beta_k} = \pi_i(1 - \pi_i)x_{ik}$; therefore, the second partial derivatives is:

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_j} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^n \left\{ x_{ij} \left(0 - \frac{\partial \pi_i}{\partial \beta_k} \right) \right\} = - \sum_{i=1}^n \pi_i(1 - \pi_i)x_{ij}x_{ik},$$

which is negative definite. Therefore, there is a unique solution $\hat{\boldsymbol{\beta}}$, the MLE for $\boldsymbol{\beta}$, because of global concavity of the log-likelihood function above. The matrix form for the second partial derivative, also known as **Hessian matrix**, is:

$$\mathbf{H} = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = - \sum_{i=1}^n \pi_i(1 - \pi_i)\mathbf{x}_i\mathbf{x}_i^T.$$

To find the ML estimates using the Newton-Raphson method, we need the second partial derivatives and an initial $\boldsymbol{\beta}_0$ as a starting point.

Recall that the variance of the Bernoulli/binomial distribution is:

- If Y_i is Bernoulli distribution with $n_i = 1$ and π_i then $V(Y_i) = \pi_i(1 - \pi_i) = \nu_i(\boldsymbol{\beta})$, and
- If Y_i is Binomial distribution $n_i > 1$ and π_i then $V(Y_i) = n_i\pi_i(1 - \pi_i) = \nu_i(\boldsymbol{\beta})$.

The vector/matrix notation. The logistic model can be written in matrix form as:

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} \text{logit}(\pi_1) \\ \vdots \\ \text{logit}(\pi_n) \end{pmatrix} = \mathbf{X}\boldsymbol{\beta},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_n \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}.$$

Also, define the vectors

$$\exp(\mathbf{X}\boldsymbol{\beta}) = \begin{pmatrix} \exp(\mathbf{x}_1^T \boldsymbol{\beta}) \\ \vdots \\ \exp(\mathbf{x}_n^T \boldsymbol{\beta}) \end{pmatrix} \Rightarrow \log \{\mathbf{1} + \exp(\mathbf{X}\boldsymbol{\beta})\} = \begin{pmatrix} \log(1 + \exp(\mathbf{x}_1^T \boldsymbol{\beta})) \\ \vdots \\ \log(1 + \exp(\mathbf{x}_n^T \boldsymbol{\beta})) \end{pmatrix},$$

where operations are performed element-wise. Then the log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^n \log(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}) = \mathbf{Y}^T \mathbf{X} \boldsymbol{\beta} - \mathbf{n}^T \log \{\mathbf{1} + \exp(\mathbf{X}\boldsymbol{\beta})\},$$

where $\mathbf{n} = \mathbf{1}$ is a vector of ones. The score function is

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0} \quad \text{where } \boldsymbol{\mu} = \mathbf{E}(\mathbf{Y}).$$

Also, we have

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = - \sum_{i=1}^n \pi_i (1 - \pi_i) x_{ij} x_{ik} = - \sum_{i=1}^n \nu_i(\boldsymbol{\beta}) x_{ij} x_{ik} \iff \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = - \sum_{i=1}^n \pi_i (1 - \pi_i) \mathbf{x}_i \mathbf{x}_i^T.$$

If we define the $n \times n$ diagonal matrix

$$\mathbf{D}(\boldsymbol{\beta}) = \text{diag} \{ \nu_1(\boldsymbol{\beta}), \nu_2(\boldsymbol{\beta}), \dots, \nu_n(\boldsymbol{\beta}) \} = \begin{pmatrix} \nu_1(\boldsymbol{\beta}) & 0 & \dots & 0 \\ 0 & \nu_2(\boldsymbol{\beta}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \nu_n(\boldsymbol{\beta}) \end{pmatrix},$$

then it is easy to show that

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = - \mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}) \mathbf{X} = - \sum_{i=1}^n \pi_i (1 - \pi_i) x_{ij} x_{ik}.$$

This is not a function of Y_i , so the *observed* and *expected* information matrix are identical, i.e. it is important to recognize that for the logistic regression model for canonical link function,

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{E} \left[- \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = \mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}) \mathbf{X} = - \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}. \quad (3.2)$$

Here the NR and Fisher Scoring methods are **equivalent**. In particular, the NR method iterates via

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \left[-\frac{\partial^2 \ell(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}_t \partial \boldsymbol{\beta}_t'} \right]^{-1} \frac{\partial \ell(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}_t} = \boldsymbol{\beta}_t + (\mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}_t) \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu}), \quad \text{for } t = 0, 1, \dots$$

until convergence to the MLE $\hat{\boldsymbol{\beta}}$.

Remark 3.1 (Line search procedure): The NR method is iterative and we use the following equation by starting at an initial guess to improve the estimate of $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \mathbf{H}^{-1} \frac{\partial \ell(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}_t} \quad \text{for } t = 0, 1, \dots \quad (3.3)$$

where the gradient $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_t}$ and the Hessian matrix, $\mathbf{H} = \mathbf{H}(\hat{\boldsymbol{\beta}}_t)$ are both evaluated at $\hat{\boldsymbol{\beta}}_t$. The vector $\delta \boldsymbol{\beta} = -\mathbf{H}^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is called the full Newton step. We use the line search procedure to guarantee the convergence of the NR iterations. While updating $\boldsymbol{\beta}_t$ with the amount $\delta \boldsymbol{\beta}$, if the new log-likelihood at $\boldsymbol{\beta}_{t+1}$ is smaller than the old log-likelihood at $\boldsymbol{\beta}_t$, we update the $\boldsymbol{\beta}_t$ by the $\frac{\delta \boldsymbol{\beta}}{2}$. This is called a line search procedure whereby we repeat iterations with half the previous step until the new log-likelihood value at $\boldsymbol{\beta}_{t+1}$ is not lower than the log-likelihood value at $\boldsymbol{\beta}_t$.

Remark 3.2: I will note that the observed information matrix $\mathbf{I}(\boldsymbol{\beta})$ is independent of $\mathbf{Y}$ for logistic regression with the **logit link** (canonical link function $\eta_i = \mu_i$), thus the observed and expected information matrices are **identical**, but not for other binomial response models, such as **probit** regression. Thus, for other models there is a difference between NR and Fisher Scoring.

3.2.2 Distribution for MLE

By the theory of maximum likelihood, $\hat{\boldsymbol{\beta}}$ has asymptotically a normal distribution:

$$\hat{\boldsymbol{\beta}} \sim AN\left(\boldsymbol{\beta}, \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})\right), \quad \text{with} \quad \mathbf{I}(\hat{\boldsymbol{\beta}}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T},$$

where $\mathbf{I}(\hat{\boldsymbol{\beta}})$ is the observed information matrix and is estimated by $\mathbf{I}(\hat{\boldsymbol{\beta}}) \approx \mathbf{X}^T \mathbf{W}(\hat{\boldsymbol{\beta}}) \mathbf{X}$. Thus, the asymptotic variance-covariance of $\hat{\boldsymbol{\beta}}$ is the inverse of the observed information matrix.

3.3 Logistic Regression for Multi-level Response

In a study, when the dependent variable has more than two categories (known as categorical Polytomous dependent variables), and we are interested in modelling this type of variable, then we generalize the binary/dichotomous logistic regression to a multinomial logistic regression. We can also model the polytomous dependent variable using the log linear model for the multiway contingency table when all the predictors are discrete; however, this methodology has two main disadvantages:

- The multinomial logistic model describes just the conditional distribution of the dependent variable given all the covariates, whereas the log linear model describes the joint distribution for all the variables, and has a large number of parameters (most of which are not of interest).
- When dealing with a set of categorical variables and one variable is obviously the response variable while all others are the covariates, the multinomial logistic model is preferable to the log linear model, which is used to describe the pattern of data in a contingency table and is therefore more complicated to interpret.

Prior to implementing multinomial logistic regression (MNLr) (also known as a *baseline-category logit model*), we need to consider the sample size and any outlying cases, whether

the dependent variable is non-metric (nominal/ordinal), and whether the independent variables are metric or dichotomous. We should also ensure that multicollinearity is evaluated. MNL is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable.

These polytomous response models can be classified into two distinct types: ordered (proportional odds and cumulative logit) [33] and unordered (sequential and multinomial) [18]. Of the multinomial model, there are three types: generalized logit (GL) model, conditional logit (CL) model and the mixed logit (ML) model. All three types of multinomial models assume that data are case-specific, that independence exists among the dependent variables, and that errors are independently and identically distributed. Note, however, that normality, linearity or homoscedasticity are not assumed.

- (1) **Generalized Logit Models** consist of a combination of several binary logits estimated at the same time. The predictors, which are the characteristics of a subject, are constant over the alternatives. The probability that subject i chooses alternative j is

$$\pi_{ij} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}{\sum_{\ell=1}^c \exp(\mathbf{x}_i^T \boldsymbol{\beta}_\ell)},$$

where $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_c$ are c vectors of unknown regression parameters (each of which is different, even though $\mathbf{x}_i$ is constant across alternatives). Since $\sum_{j=1}^c \pi_{ij} = 1$, the c sets of parameters are not unique. To fit the GL model, we set the reference category parameters to zero (i.e. $\boldsymbol{\beta}_c = \mathbf{0}$), and then we only have to estimate $c - 1$ sets of regression parameters [18].

- (2) **Conditional Logit Models** are a modified version of the GL model in that they assign a survival/failure time for each alternative. Choice represents the failure and has a value of 1 for most preferred choice, 2 otherwise. For an example of this, refer to [18] under Example 2: Travel Choice Data. In the CL model, the explanatory variables $\mathbf{z}_j$ represent a vector of characteristics of the j^{th} alternative, which assume different values for each

alternative, where the impact of a unit of $\mathbf{z}_j$ is assumed to be constant across alternatives. The probability that individual i chooses alternative j is

$$\pi_{ij} = \frac{\exp(Z_{ij}^T \boldsymbol{\theta})}{\sum_{\ell=1}^c \exp(Z_{i\ell}^T \boldsymbol{\theta})},$$

where $\boldsymbol{\theta}$ is a single vector of regression parameters.

- (3) **Mixed Logit Models** include both characteristics of the individual and the alternatives. We calculate the choice probabilities using

$$\pi_{ij} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_j + Z_{ij}^T \boldsymbol{\theta})}{\sum_{\ell=1}^c \exp(\mathbf{x}_i^T \boldsymbol{\beta}_\ell + Z_{i\ell}^T \boldsymbol{\theta})},$$

where $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_g$ and $\boldsymbol{\beta}_c \equiv \mathbf{0}$ are the alternative-specific parameters, and $\boldsymbol{\theta}$ is the set of global parameters.

For the purposes of this thesis, Subsection 3.3.1 will focus on the generalized logit model, and will present a model for nominal responses that uses a separate binary logit model for each pair of response categories.

3.3.1 Nominal Responses: Baseline-Category Logit Models

For a nominal-scale response variable Y with c categories, multicategory (also called polytomous) logit models for nominal response variables simultaneously describe log odds for all cC_2 pairs of categories. However, this kind of specification is redundant as only $g = c - 1$ of the $\pi_j(\mathbf{x}_i) = \pi_{ij}$ are independent [8]. For a set of explanatory/independent variables $\mathbf{x} = (x_1, \dots, x_p)^T$, let

$$\pi_{ij} = \pi_j(\mathbf{x}_i) = P(Y_i = j | \mathbf{x}_i), \quad j = 1, \dots, c \text{ and } \sum_{j=1}^c \pi_j(\mathbf{x}_i) = 1. \quad (3.4)$$

In Section 3.2 we defined odds for a binary response as $\frac{P(\text{success})}{P(\text{failure})}$ [7]. More generally, for a multinomial response, we can define odds to be a comparison of any pair of response categories. For example, $\frac{\pi_{ij}}{\pi_{ij'}}$ is the odds of category j relative to j' . The *generalized logit* model designates one category as a reference level and then pairs each other response category with this reference category. Usually, the first or the last category is chosen to serve as such a reference category. Of course, for nominal responses, the first or last category is not well-defined since the categories are exchangeable. Thus, the selection of the reference level is arbitrary and is typically based on convenience. For multinomial responses, we have more than two response levels and as such cannot define odds or log odds of response as in the binary case. However, upon selecting the reference level, say the last level c , we can define the odds (log odds) of response in the j -th category as compared to the c -th response category by $\frac{\pi_{ij}}{\pi_{ic}} \left(\log \left(\frac{\pi_{ij}}{\pi_{ic}} \right) \right)$ ($1 \leq j \leq g$). Note that since $\pi_j(\mathbf{x}_i) + \pi_c(\mathbf{x}_i) \neq 1$, $\frac{\pi_{ij}}{\pi_{ic}} \left(\log \left(\frac{\pi_{ij}}{\pi_{ic}} \right) \right)$ is not odds (log odds) in the usual sense. However, we have

$$\log \frac{\pi_{ij}}{\pi_{ic}} = \log \frac{\pi_{ij}/(\pi_{ij} + \pi_{ic})}{\pi_{ic}/(\pi_{ij} + \pi_{ic})} = \log \frac{\pi_{ij}/(\pi_{ij} + \pi_{ic})}{1 - \pi_{ij}/(\pi_{ij} + \pi_{ic})} = \text{logit} \left(\frac{\pi_{ij}}{\pi_{ij} + \pi_{ic}} \right).$$

Thus, $\log \left(\frac{\pi_{ij}}{\pi_{ic}} \right)$ has the usual log odds interpretation if we limit our interest to the two levels ℓ and c , which explains the name: generalized logit model. We model the log odds of responses for each pair of categories by relating a set of explanatory variables to each log-odds as follows:

$$\begin{aligned} \log \left(\frac{\pi_{ij}}{\pi_{ic}} \right) &= \log \left(\frac{\pi_j(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i)} \right) = \log \left(\frac{\pi_{ij}}{1 - \sum_{j=1}^g \pi_{ij}} \right) = \beta_{0j} + \beta_{1j}x_{i1} + \cdots + \beta_{pj}x_{ip} \\ &= \sum_{k=0}^p x_{ik}\beta_{kj} = \beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j \\ &= \eta_{ij} \end{aligned} \tag{3.5}$$

for $j = 1, \dots, g$, where $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \dots, \beta_{pj})^T$ and

* for each subject, y_{ij} represents the observed counts of the j -th value of Y_{ij} ,

- * π_{ij} is the probability of observing the j -th value of the dependent variable for any given observation in the i -th subject,
- * β_{kj} contains the parameter estimate for the k -th covariate and the j -th value of the dependent variable.

Note that the second subscript on the β parameters corresponds to the response category, which allows each response's log-odds to relate to the explanatory variables in a different way. Since

$$\log \frac{\pi_{ij}}{\pi_{i\ell}} = \log \frac{\pi_{ij}}{\pi_{ic}} - \log \frac{\pi_{i\ell}}{\pi_{ic}},$$

these g logits determine the parameters for any other pairs of the response categories. The probabilities for each individual category can also be found in terms of the model. We can re-write equation (3.5) as:

$$\pi_{ij} = \pi_{ic} \exp(\eta_{ij}) = \pi_{ic} \exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)$$

using properties of logarithms. Noting that $\sum_{j=1}^c \pi_{ij} = 1$, we have

$$\pi_{i1} + \pi_{i2} + \cdots + \pi_{ic} = 1 = \pi_{ic} \exp(\beta_{01} + \mathbf{x}_i^T \boldsymbol{\beta}_1) + \pi_{ic} \exp(\beta_{02} + \mathbf{x}_i^T \boldsymbol{\beta}_2) + \cdots + \pi_{ic}.$$

By factoring out the reference probability π_{ic} in each term, we obtain an expression for π_{ic} :

$$\pi_{ic} = \frac{1}{1 + \sum_{j=1}^g \exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)} = \frac{1}{1 + \sum_{j=1}^g \exp(\eta_{ij})}. \quad (3.6)$$

This leads to a general expression for π_{ij} for $j = 1, \dots, c-1$.

$$\pi_{ij} = \frac{\exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{j=1}^g \exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)} = \frac{\exp(\eta_{ij})}{1 + \sum_{j=1}^g \exp(\eta_{ij})} \quad (3.7)$$

3.3.2 Estimation of Model Parameters

Parameters for the model are estimated using ML. For a sample of observations, Y_i , that denote the category response and corresponding explanatory variables $x_{i1}, \dots, x_{ip}$ for $i = 1, \dots, r$, the likelihood function (the joint density function) is simply the product of r multinomial distributions with probabilities as given by Equations (3.6) and (3.7). The joint density is given by

$$f(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^r \left[\frac{n_i!}{\prod_{j=1}^c y_{ij}!} \prod_{j=1}^c \pi_{ij}^{y_{ij}} \right],$$

where

- * $\mathbf{Y}$ is a response matrix with r rows (one for each subject) and $g = c - 1$ columns,
- * y_{ij} represents the observed counts of the j -th value of Y_{ij} ,
- * $\boldsymbol{\pi}$ is a matrix of the same dimensions as $\mathbf{Y}$, where each element π_{ij} is the probability of observing the j -th value of the dependent variable for any given observation in the i -th subject,
- * $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_r^T \end{pmatrix} = \begin{pmatrix} x_{10} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i0} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{r0} & \cdots & x_{rk} & \cdots & x_{rp} \end{pmatrix}$ is the design matrix of independent variables, with r rows and $p + 1$ columns where p is the number of independent variables and the first element of each row, $x_{i0} = 1$, corresponding to the intercept term,
- * $\boldsymbol{\beta} = \begin{pmatrix} \beta_{01} & \cdots & \beta_{0j} & \cdots & \beta_{0g} \\ \vdots & \ddots & \vdots & & \vdots \\ \beta_{k1} & \cdots & \beta_{kj} & \cdots & \beta_{kg} \\ \vdots & & \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pj} & \cdots & \beta_{pg} \end{pmatrix}$ is a matrix with $p + 1$ rows and g columns, such that each element β_{kj} contains the parameter estimate for the k -th covariate and the j -th value of the dependent variable,
- * $\mathbf{n}$ is the column vector that contains elements n_i , which represent the number of observations in subject i , such that $\sum_{i=1}^r n_i = n$, the total sample size.

Thus, the likelihood function for the multinomial logistic regression model (without the constant term) is given by

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^r \prod_{j=1}^c \pi_{ij}^{y_{ij}},$$

which can be rewritten as

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^r \prod_{j=1}^g \pi_{ij}^{y_{ij}} \pi_{ic}^{n_i - \sum_{j=1}^g y_{ij}} = \prod_{i=1}^r \prod_{j=1}^g \left(\frac{\pi_{ij}}{\pi_{ic}} \right)^{y_{ij}} \pi_{ic}^{n_i}.$$

Now, substitute for π_{ij} and π_{ic} using Equations (3.6) and (3.7). Then we have the likelihood

$$\begin{aligned} L(\boldsymbol{\beta}|\mathbf{y}) &= \prod_{i=1}^r \prod_{j=1}^g (\exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j))^{y_{ij}} \left(\frac{1}{1 + \sum_{j=1}^g \exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)} \right)^{n_i} \\ &= \prod_{i=1}^r \prod_{j=1}^g (\exp(\eta_{ij}))^{y_{ij}} \left(1 + \sum_{j=1}^g \exp(\eta_{ij}) \right)^{-n_i}. \end{aligned}$$

The corresponding log-likelihood is given by

$$\begin{aligned} \ell(\boldsymbol{\beta}|\mathbf{y}) &= \sum_{i=1}^r \sum_{j=1}^g \{y_{ij} \ln(\exp(\eta_{ij}))\} - \sum_{i=1}^r \left\{ n_i \ln \left(1 + \sum_{j=1}^g \exp(\eta_{ij}) \right) \right\} \\ &= \sum_{i=1}^r \sum_{j=1}^g \left(y_{ij} \sum_{k=0}^p x_{ik} \beta_{kj} \right) - \sum_{i=1}^r \left\{ n_i \ln \left(1 + \sum_{j=1}^g \exp \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right) \right) \right\}. \end{aligned}$$

The sufficient statistic for β_{kj} is $\sum_{i=1}^r x_{ik} y_{ij}$, $j = 1, \dots, g$, $k = 1, \dots, p$. The sufficient statistic for β_{0j} is $\sum_{i=1}^r y_{ij} = \sum_{i=1}^r x_{i0} y_{ij}$ for $x_{i0} = 1$; this is the total number of outcomes in category j .

✱ There are $q = (p + 1).g$ parameters to be estimated. Let $\boldsymbol{\Theta}$ be the entire parameter set denoted by:

$$\boldsymbol{\Theta} = \{(\beta_{01}, \dots, \beta_{p1}), (\beta_{02}, \dots, \beta_{p2}), \dots, (\beta_{0j}, \dots, \beta_{pj}), \dots, (\beta_{0g}, \dots, \beta_{pg})\}.$$

The above set is organized with $p + 1$ coefficients for first response function, then $p + 1$ coefficients for second response function, etc.

To maximize $\ell(\boldsymbol{\beta})$, we compute the score function by taking the first partial derivatives and set them equal to zero:

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{01}} & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{02}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{0j}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{0g}} \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{11}} & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{12}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{1j}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{1g}} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{k1}} & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{k2}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{kj}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{kg}} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{p1}} & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{p2}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{pj}} & \cdots & \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{pg}} \end{pmatrix} = \mathbf{0}.$$

So we are solving a system of $q = g.(p + 1)$ non-linear equations and solve for each β_{kj} . Let us now compute $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{kj}}$ where β_{kj} is an ij -th element of $\boldsymbol{\beta}$. Thus each of the partial derivatives in $S(\boldsymbol{\beta})$ will have the same form:

$$S(\beta_{kj}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{kj}} = \sum_{i=1}^r \sum_{j=1}^g y_{ij} \frac{\partial \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right)}{\partial \beta_{kj}} - \sum_{i=1}^r n_i \frac{\partial \left\{ \ln \left(1 + \sum_{j=1}^g \exp \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right) \right) \right\}}{\partial \beta_{kj}},$$

where

$$\frac{\partial(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{\partial \beta_{kj}} = \frac{\partial(\beta_{0j} + x_{i1}\beta_{1j} + \cdots + x_{ik}\beta_{kj} + \cdots + x_{ip}\beta_{pj})}{\partial \beta_{kj}} = x_{ik} \quad \text{where } x_{i0} = 1,$$

and

$$\begin{aligned} \frac{\partial \left\{ \ln \left(1 + \sum_{j=1}^g \exp \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right) \right) \right\}}{\partial \beta_{kj}} &= \frac{1}{1 + \sum_{j=1}^g \exp \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right)} \frac{\partial}{\partial \beta_{kj}} \left(1 + \sum_{j=1}^g \exp \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right) \right) \\ &= \frac{\exp \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right)}{1 + \sum_{j=1}^g \exp \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right)} \frac{\partial}{\partial \beta_{kj}} \left(\sum_{k=0}^p x_{ik} \beta_{kj} \right) \\ &= \pi_{ij} x_{ik}. \end{aligned}$$

Therefore, we have the score function

$$S(\beta_{kj}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{kj}} = \sum_{i=1}^r (y_{ij} x_{ik} - n_i \pi_{ij} x_{ik}) = \sum_{i=1}^r (y_{ij} - n_i \pi_{ij}) x_{ik} = s_{kj}. \quad (3.8)$$

For each β_{kj} , we need to differentiate equation (3.8) with respect to every other β_{kj} . This way, we can form the matrix of second partial derivatives of order $q = g \cdot (p + 1)$ as:

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{k'j'}} &= \frac{\partial}{\partial \beta_{k'j'}} \sum_{i=1}^r (y_{ij} - n_i \pi_{ij}) x_{ik} = - \frac{\partial}{\partial \beta_{k'j'}} \sum_{i=1}^r n_i \pi_{ij} x_{ik} \\ &= - \sum_{i=1}^r n_i x_{ik} \frac{\partial}{\partial \beta_{k'j'}} \left(\frac{\exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{j=1}^g \exp(\beta_{0j} + \mathbf{x}_i^T \boldsymbol{\beta}_j)} \right) \\ &= - \sum_{i=1}^r n_i x_{ik} \frac{\partial}{\partial \beta_{k'j'}} \left(\frac{\exp(\eta_{ij})}{1 + \sum_{j=1}^g \exp(\eta_{ij})} \right). \end{aligned}$$

Using Cauchy's rule, we have

$$\frac{\partial \exp(\eta_{ij})}{\partial \beta_{k'j'}} = \begin{cases} \frac{\partial \exp(\eta_{ij})}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \beta_{k'j'}} = \exp(\eta_{ij}) x_{ik'} & \text{for } j' = j \\ \frac{\partial \exp(\eta_{ij})}{\partial \eta_{ij'}} \frac{\partial \eta_{ij'}}{\partial \beta_{k'j'}} = \exp(\eta_{ij})(0) = 0 & \text{for } j' \neq j \end{cases},$$

and

$$\frac{\partial \left(1 + \sum_{j=1}^g \exp(\eta_{ij}) \right)}{\partial \beta_{k'j'}} = \begin{cases} \frac{\partial \left(1 + \sum_{j=1}^g \exp(\eta_{ij}) \right)}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \beta_{k'j'}} = \exp(\eta_{ij}) x_{ik'} & \text{for } j' = j \\ \frac{\partial \left(1 + \sum_{j=1}^g \exp(\eta_{ij}) \right)}{\partial \eta_{ij'}} \frac{\partial \eta_{ij'}}{\partial \beta_{k'j'}} = \exp(\eta_{ij'}) x_{ik'} & \text{for } j' \neq j \end{cases}.$$

Then using the quotient rule, the second partial derivatives are

$$\begin{aligned}
\frac{\partial}{\partial \beta_{k'j'}} \left(\frac{\exp(\eta_{ij})}{1 + \sum_{j=1}^g \exp(\eta_{ij})} \right) &= \begin{cases} \frac{\left(1 + \sum_{j=1}^g \exp(\eta_{ij})\right) \exp(\eta_{ij}) x_{ik'} - \exp(\eta_{ij}) \exp(\eta_{ij}) x_{ik'}}{\left(1 + \sum_{j=1}^g \exp(\eta_{ij})\right)^2} & \text{for } j' = j \\ \frac{0 - \exp(\eta_{ij}) \exp(\eta_{ij'}) x_{ik'}}{\left(1 + \sum_{j=1}^g \exp(\eta_{ij})\right)^2} & \text{for } j' \neq j \end{cases} \\
&= \begin{cases} \pi_{ij}(1 - \pi_{ij}) x_{ik'} & \text{for } j' = j \\ -\pi_{ij} \pi_{ij'} x_{ik'} & \text{for } j' \neq j \end{cases}.
\end{aligned}$$

Therefore the full square matrix of $q = g.(p + 1)$ order for second partial derivatives of multinomial logistic regression model looks like:

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{01} \partial \beta_{01}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{01} \partial \beta_{p1}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{01} \partial \beta_{02}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{01} \partial \beta_{p2}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{01} \partial \beta_{0g}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{01} \partial \beta_{pg}} \\ \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_{01}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_{p1}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_{02}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_{p2}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_{0g}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{11} \partial \beta_{pg}} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots \\ \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{01}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{p1}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{02}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{p2}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{0g}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{pg}} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots \\ \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{pg} \partial \beta_{01}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{pg} \partial \beta_{p1}} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{pg} \partial \beta_{02}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{pg} \partial \beta_{p2}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{pg} \partial \beta_{0g}} & \dots & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{pg} \partial \beta_{pg}} \end{pmatrix},$$

where

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{k'j'}} = \begin{cases} -\sum_{i=1}^r n_i \pi_{ij}(1 - \pi_{ij}) x_{ik} x_{ik'} & \text{for } j' = j \\ \sum_{i=1}^r n_i \pi_{ij} \pi_{ij'} x_{ik} x_{ik'} & \text{for } j' \neq j \end{cases}. \quad (3.9)$$

Because the equations in (3.9) are different based on whether or not $j' = j$, the second partial derivatives matrix in the multinomial case is somewhat different from that derived in the binomial case.

- * For the diagonal elements of the full matrix, where $j' = j$, there is a set of g square sub-matrices of order r , which are denoted as

$$\mathbf{D}_{jj} = \begin{pmatrix} n_1\pi_{1j}(1 - \pi_{1j}) & 0 & \cdots & 0 \\ 0 & n_2\pi_{2j}(1 - \pi_{2j}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_r\pi_{rj}(1 - \pi_{rj}) \end{pmatrix}, \quad (3.10)$$

where the i^{th} diagonal elements of each sub-matrix are computed using $n_i\pi_{ij}(1 - \pi_{ij})$, and the off-diagonal elements are zeros. $1 \leq j \leq g$.

- * For the off-diagonal elements of the full matrix, where $j' \neq j$, the square sub-matrices of order r are denoted as

$$\mathbf{D}_{jj'} = \begin{pmatrix} -n_1\pi_{1j}\pi_{1j'} & 0 & \cdots & 0 \\ 0 & -n_2\pi_{2j}\pi_{2j'} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -n_r\pi_{rj}\pi_{rj'} \end{pmatrix}, \quad (3.11)$$

where the i^{th} diagonal elements of each sub-matrix are computed using $-n_i\pi_{ij}\pi_{ij'}$, and the off-diagonal elements are zeros. $1 \leq j, j' \leq g$.

- * Now let us construct Hessian matrices $\mathbf{H}_{jj} = \mathbf{X}^T \mathbf{D}_{jj} \mathbf{X}$ for $j' = j$ and $\mathbf{H}_{jj'} = \mathbf{X}^T \mathbf{D}_{jj'} \mathbf{X}$ for $j' \neq j$ of order $(p + 1) \times (p + 1)$, where $1 \leq j, j' \leq g$.
- * The above Hessian matrices are negative definite matrices. Therefore, the full Hessian matrix of $q \times q$ order is:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \cdots & \mathbf{H}_{1g} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \cdots & \mathbf{H}_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{g1} & \mathbf{H}_{g2} & \cdots & \mathbf{H}_{gg} \end{pmatrix}.$$

The Hessian matrix is a negative definite matrix. Thus, there is a unique solution $\hat{\boldsymbol{\beta}}$, the MLE for $\boldsymbol{\beta}$.

- $E(Y_{ij}) = n_i \pi_{ij}$, $V(Y_{ij}) = n_i \pi_{ij}(1 - \pi_{ij})$ and $\text{Cov}(Y_{ij}, Y_{i\ell}) = -n_i \pi_{ij} \pi_{i\ell}$

The vector/matrix notation. For the multinomial logistic model, the log-likelihood is formed element-wise as

$$\ell(\boldsymbol{\beta}) = \text{tr}(\mathbf{Y}^T \mathbf{X} \boldsymbol{\beta}) - \mathbf{n}^T \log \left\{ \mathbf{1} + \sum_{j=1}^g \exp(\mathbf{X} \boldsymbol{\beta}) \right\} = \text{tr}(\mathbf{Y}^T \mathbf{X} \boldsymbol{\beta}) - \mathbf{n}^T \log \left\{ \mathbf{1} + \sum_{j=1}^g \begin{pmatrix} \exp(\eta_{1j}) \\ \vdots \\ \exp(\eta_{rj}) \end{pmatrix} \right\},$$

and the ML estimating equations take the form

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0},$$

where the mean matrix $\boldsymbol{\mu}$ of order $r \times g$ is computed by multiplying the i^{th} row of probability matrix $\boldsymbol{\pi}$ by n_i , $i = 1, \dots, r$. The matrix form for the second derivative is:

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_{j'}^T} = \begin{cases} -\sum_{i=1}^r n_i \pi_{ij} (1 - \pi_{ij}) \mathbf{x}_i \mathbf{x}_i^T & \text{for } j' = j \\ \sum_{i=1}^r n_i \pi_{ij} \pi_{ij'} \mathbf{x}_i \mathbf{x}_i^T & \text{for } j' \neq j \end{cases}.$$

Therefore to illustrate the iterative procedure of Newton-Raphson (NR) for the multinomial logistic regression model, we need to convert:

- (1) the $(p+1) \times g$ coefficient matrix $\boldsymbol{\beta}$ into a vector $\boldsymbol{\beta}$ of length $q = (p+1).g$, and
- (2) the $(p+1) \times g$ matrix score function $S(\boldsymbol{\beta})$ into a vector $S(\boldsymbol{\beta})$ of length q .

Then the iterative expression below for each coefficient matrix $\boldsymbol{\beta}$:

$$\text{vec}(\boldsymbol{\beta}_{t+1}) = \text{vec}(\boldsymbol{\beta}_t) - \left[-\frac{\partial^2 \ell(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}_t \partial \boldsymbol{\beta}_t^T} \right]^{-1} \frac{\partial \ell(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}_t} = \text{vec}(\boldsymbol{\beta}_t) + \mathbf{H}^{-1} \text{vec}(\mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu})) \quad (3.12)$$

for $t = 0, 1, \dots$ until convergence to the MLE $\hat{\boldsymbol{\beta}}$.

3.4 Multinomial Logit Model as Multivariate GLM

Consider a random variable Y that may take one of the category j . Let $\pi_j = P(Y = j|\mathbf{x})$ be the response probability associated with $(p+1)$ -dimensional vector, $\mathbf{x}$, of covariates, where $x_0 = 1$, and the linear predictor $\eta_j = \mathbf{x}^T \boldsymbol{\beta}_j$. We may write the g equations for nominal logit model with reference category c in matrix form by

$$\underbrace{\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_g \end{pmatrix}}_{g \times 1} = \underbrace{\begin{pmatrix} \log\left(\frac{\pi_1}{\pi_c}\right) \\ \vdots \\ \log\left(\frac{\pi_g}{\pi_c}\right) \end{pmatrix}}_{g \times 1} = \underbrace{\begin{pmatrix} \mathbf{x}^T & \cdots & \mathbf{0}^T \\ \vdots & \ddots & \vdots \\ \mathbf{0}^T & \cdots & \mathbf{x}^T \end{pmatrix}}_{g \times q = g \times (p+1)g} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_g \end{pmatrix}}_{q \times 1} = \underbrace{\begin{pmatrix} \mathbf{x}^T \boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{x}^T \boldsymbol{\beta}_g \end{pmatrix}}_{g \times 1}. \quad (3.13)$$

For the j^{th} category, $\log\left(\frac{\pi_j}{\pi_c}\right)$ has the form

$$\eta_j = \mathbf{x}^T \boldsymbol{\beta}_j = (0, \dots, 0, \mathbf{x}^T, 0, \dots, 0) \boldsymbol{\beta} = \mathbf{x}_j^T \boldsymbol{\beta},$$

where $\mathbf{x}_j^T = (0, \dots, 0, \mathbf{x}^T, 0, \dots, 0)$ is the corresponding design vector and the coefficient column vector $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_g^T)$ of length $q = (p+1)g$, which are obtained by stacking g columns in the matrix $\boldsymbol{\beta}$ on top of one another. Thus, for **one** observation, the nominal logit model has the general form of Eq. (3.13):

$$\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\pi}) = (g_1(\boldsymbol{\pi}), \dots, g_g(\boldsymbol{\pi})) = \left(\log\left(\frac{\pi_1}{\pi_c}\right), \dots, \log\left(\frac{\pi_g}{\pi_c}\right) \right) = \mathbf{X} \boldsymbol{\beta} \quad \text{or}$$

$$\boldsymbol{\pi} = \mathbf{h}(\mathbf{X} \boldsymbol{\beta}) = (h_1(\boldsymbol{\eta}), \dots, h_g(\boldsymbol{\eta})) = \left(\frac{\exp(\eta_1)}{1 + \sum_{j=1}^g \exp(\eta_j)}, \dots, \frac{\exp(\eta_g)}{1 + \sum_{j=1}^g \exp(\eta_j)} \right),$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)^T$ is the vector of the response probabilities, $\mathbf{X}$ is a design matrix of order $g \times q$ that corresponds to the whole parameter vector $\boldsymbol{\beta}$, and $\mathbf{g}$ and $\mathbf{h} = \mathbf{g}^{-1}$ are the vector-valued link and response functions, respectively, in generalized linear models.

For the given data $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \dots, r$, the model for the i^{th} observation is

$$\boldsymbol{\eta}_i = \mathbf{g}(\boldsymbol{\pi}_i) = \mathbf{X}_i \boldsymbol{\beta} \quad \text{or} \quad \boldsymbol{\pi}_i = \boldsymbol{\pi}(\mathbf{x}_i) = \mathbf{h}(\mathbf{X}_i \boldsymbol{\beta}),$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{ig})^T$, $\pi_{ij} = P(Y_i = j | \mathbf{x}_i) = \frac{\exp(\eta_{ij})}{1 + \sum_{\ell=1}^g \exp(\eta_{i\ell})}$, and $\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_i^T & \mathbf{0}^T & \dots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{x}_i^T & \dots & \mathbf{0}^T \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \dots & \mathbf{x}_i^T \end{pmatrix}$ is a matrix of order $g \times q$ composed of the $(p+1)$ -dimensional vector, $\mathbf{x}_i$, of covariates, where $x_{i0} = 1$. For more details, see Section 8.1.5 in the 3rd edition of Categorical Data Analysis by Alan Agresti [1].

3.4.1 Maximum Likelihood Estimation

The multinomial distribution has the form of a multivariate exponential family (EF), see Appendix B. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{ig})^T \sim MN(n_i, \boldsymbol{\pi}_i)$, $i = 1, \dots, r$, denote the multinomial distribution with $c = g + 1$ categories. Then the probability mass function is

$$\begin{aligned} f(\mathbf{y}_i) &= \frac{n_i!}{\prod_{j=1}^c y_{ij}!} \prod_{j=1}^c \pi_{ij}^{y_{ij}} = \frac{n_i!}{y_{i1}! \dots y_{ig}! (n_i - \sum_{j=1}^g y_{ij})!} \pi_{i1}^{y_{i1}} \dots \pi_{ig}^{y_{ig}} (1 - \sum_{j=1}^g \pi_{ij})^{(n_i - \sum_{j=1}^g y_{ij})} \\ &= \exp \left\{ \mathbf{y}_i^T \boldsymbol{\theta}_i + n_i \log(\pi_{ic}) + \log(c_i) \right\} \\ &= \exp \left\{ n_i \mathbf{p}_i^T \boldsymbol{\theta}_i + n_i \log(\pi_{ic}) + \log(c_i) \right\}, \end{aligned} \tag{3.14}$$

where the canonical parameter vector is $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{ig})^T$, $\theta_{ij} = \log \left(\frac{\pi_{ij}}{\pi_{ic}} \right)$, $\pi_{ic} = 1 - \sum_{j=1}^g \pi_{ij}$, the dispersion parameter is $\frac{1}{n_i}$, $c_i = \frac{n_i!}{\prod_{j=1}^c y_{ij}!}$, and we consider $\mathbf{p}_i = \frac{1}{n_i} \mathbf{y}_i$ the scaled multinomials or proportions. Thus the likelihood is formed as

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^r f(\mathbf{y}_i) = \prod_{i=1}^r \frac{n_i!}{\prod_{j=1}^c y_{ij}!} \prod_{j=1}^c \pi_{ij}^{y_{ij}} \tag{3.15}$$

and the log-likelihood $\ell(\boldsymbol{\beta}) = \sum_{i=1}^r \ell_i(\boldsymbol{\pi}_i)$ with

$$\begin{aligned}
\ell_i(\boldsymbol{\pi}_i) &= \sum_{j=1}^g y_{ij} \theta_{ij} + n_i \log(\pi_{ic}) + \log(c_i) \\
&= \mathbf{y}_i^T \boldsymbol{\theta}_i + n_i \log(\pi_{ic}) + \log(c_i).
\end{aligned} \tag{3.16}$$

Similarly to the logistic regression for binary response/classification in Section 3.2, the first and second partial derivative for the log-likelihood are the score function $S(\boldsymbol{\beta})$ and the expected information or Fisher matrix $F(\boldsymbol{\beta})$.

The score function $S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ for the model $\boldsymbol{\pi}_i = \mathbf{h}(\mathbf{X}_i \boldsymbol{\beta})$ has the form

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^r \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}) (\mathbf{y}_i - n_i \boldsymbol{\pi}_i), \tag{3.17}$$

where $\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta}$ is the row g -dimensional vector for the linear predictor and the $g \times g$ derivatives matrix $\mathbf{D}_i(\boldsymbol{\beta}) = \frac{\partial \mathbf{h}(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\eta}} = \left(\frac{\partial \mathbf{g}(\boldsymbol{\pi}_i)}{\partial \boldsymbol{\pi}} \right)^{-1}$, which is not a symmetric matrix with entries $\frac{\partial h_j(\boldsymbol{\eta}_i)}{\partial \eta_\ell}$. The variance-covariance matrix $\Sigma_i(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ can be computed by the multinomial distribution and has the form

$$\Sigma_i(\boldsymbol{\beta}) = \frac{1}{n_i} \left(\text{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^T \right).$$

In closed matrix notation/form the score is obtained as

$$S(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}) \Sigma(\boldsymbol{\beta})^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{D}(\boldsymbol{\beta})^{-T} (\mathbf{y} - \boldsymbol{\mu}), \tag{3.18}$$

where $\mathbf{X}^T = (\mathbf{X}_1^T, \dots, \mathbf{X}_r^T)$ of order $q \times rg$ or $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_r \end{pmatrix}$ of order $rg \times q$ is a design matrix, $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_r^T)^T$ and $\boldsymbol{\mu} = (n_1 \boldsymbol{\pi}_1^T, \dots, n_r \boldsymbol{\pi}_r^T)^T$ are $rg \times 1$ response and mean vectors. Each vector is obtained by stacking r rows in the respective response or probability matrix on top of one another; this is done by appending each of the additional rows below the first. $\mathbf{D}(\boldsymbol{\beta})$ and $\Sigma(\boldsymbol{\beta})$ are block-diagonal matrices of order $rg \times rg$ with blocks $\mathbf{D}_i(\boldsymbol{\beta})$, $\Sigma_i(\boldsymbol{\beta})$, respectively, and the weight matrix $\mathbf{W}(\boldsymbol{\beta}) = \mathbf{D}(\boldsymbol{\beta}) \Sigma(\boldsymbol{\beta})^{-1} \mathbf{D}(\boldsymbol{\beta})^T$ is a block-diagonal matrix with blocks $\mathbf{W}_i(\boldsymbol{\beta}) = \mathbf{D}_i(\boldsymbol{\beta}) \Sigma_i(\boldsymbol{\beta})^{-1} \mathbf{D}_i(\boldsymbol{\beta})^T$. These weights can be obtained by $\mathbf{W}_i(\boldsymbol{\beta}) = \left(\frac{\partial \mathbf{g}(\boldsymbol{\pi}_i)}{\partial \boldsymbol{\pi}^T} \Sigma_i(\boldsymbol{\beta}) \frac{\partial \mathbf{g}(\boldsymbol{\pi}_i)}{\partial \boldsymbol{\pi}} \right)^{-1}$, which

is an approximation to the inverse of the covariance of $\mathbf{g}(\mathbf{p}_i)$ when the model can be applied.

The expected Fisher information $\mathbf{F}(\boldsymbol{\beta}) = -\mathbb{E} \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) = \text{Cov}(S(\boldsymbol{\beta}))$ has the form

$$\mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^r \mathbf{X}_i^T \mathbf{W}_i(\boldsymbol{\beta}) \mathbf{X}_i = \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}. \quad (3.19)$$

The formula for updating the MLE for $\boldsymbol{\beta}$ in binomial cases also holds for multinomial cases to update $\boldsymbol{\beta}$. The NR method iterates via

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \mathbf{F}(\boldsymbol{\beta}_t)^{-1} S(\boldsymbol{\beta}_t), \quad \text{for } t = 0, 1, \dots \quad (3.20)$$

until convergence to the MLE $\hat{\boldsymbol{\beta}}$.

For the logit model, which corresponds to the **canonical** link, the score function and the Fisher matrix have simpler forms denoted as

$$S(\boldsymbol{\beta}) = \sum_{i=1}^r n_i \mathbf{X}_i^T (\mathbf{p}_i - \boldsymbol{\pi}_i) \quad \text{and} \quad \mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^r n_i^2 \mathbf{X}_i^T \Sigma_i(\boldsymbol{\beta}) \mathbf{X}_i.$$

3.4.2 Distribution for Multinomial Logit MLE

Under regularity conditions of the maximum likelihood theorem, the ML estimate $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with

$$\hat{\boldsymbol{\beta}} \sim N_q \left(\boldsymbol{\beta}, \mathbf{F}(\hat{\boldsymbol{\beta}})^{-1} \right) \quad \text{as } N = \sum_{i=1}^r n_i \rightarrow \infty; \quad (3.21)$$

for details see Fahrmeir and Kaufmann (1985) [61]. The score function and Fisher matrix have the same forms as in univariate Generalized Linear Models (GLMs), namely, $S(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}) \Sigma(\boldsymbol{\beta})^{-1} (\mathbf{y} - \boldsymbol{\mu})$ and $\mathbf{F}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}$. But for multicategorical responses, the design matrix is composed of matrices for single observations, and the weight matrix $\mathbf{W}(\boldsymbol{\beta})$ as well as the matrix of derivatives $\mathbf{D}(\boldsymbol{\beta})$ are block-diagonal matrices in contrast to the univariate models, where $\mathbf{W}(\boldsymbol{\beta})$ and $\mathbf{D}(\boldsymbol{\beta})$ are diagonal matrices.

3.5 Multinomial Simulation Results

To illustrate the Newton-Raphson technique on the multinomial logit model, simulations were conducted using a fictitious Canadian political election study, which investigates voter intent based on gender (one covariate: male vs. female) for the following parties: 1=Conservative party (CP), 2=Liberal party (LP), and 3=Other [24]. We create two different scenarios using two sample sizes $N = 1200$ and $N = 2400$, where the number of males is 716 and females is 484, and the number of males is 1290 and females is 1110, respectively. To generate the observed data y_{ij} for these two scenarios and three categories from a multinomial logit model, the following steps were followed:

* the true parameter matrix $\beta =$

	β_1	β_2
Intercept	β_{01}	β_{02}
Gender	β_{11}	β_{12}

 $=$

	β_1	β_2
Intercept	-0.157	0.593
Gender	-0.885	-2.333

where the parameters β_1 and β_2 represent the coefficients for CP and LP categories, respectively. As the third category is the base reference category, a parameter vector is not needed [1]. The above parameter values were generated from standard normal.

* For this study, the 2×2 covariate matrix is $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, where the first column represents the intercept and the second column represents the gender. The gender covariate is recorded as 1=female, 0=male. Since we are only studying the effect of gender on voter intent, the number of parameters for j^{th} category is $p = 2$ (i.e. $j = 1, 2$).

* Using the covariate and parameter matrices, a linear combination matrix η of order $r \times c$ is computed. In this study, the dimensions for the η matrix follow the two scenarios, whereby dimensions are 2×3 . The first and second columns represent the linear combination for the CP and LP categories, respectively, and the third column is zeros which represents the reference category, "Other".

- * The probability matrix $\boldsymbol{\pi}$ of order $r \times c$ is computed from the $\boldsymbol{\eta}$ matrix entries by using the multinomial logit probabilities definitions found in Eq. (3.6) and Eq. (3.7). In this study, the probability matrix also has the two scenarios whereby dimensions are 2×3 . The first and second columns in the probability matrix represent the probabilities of “CP” and “LP” categories at each gender, respectively. The third column represents the probability of the third category “Other” at each gender. There are two sets of probability values for each category “CP”, “LP” and “Other” due to gender, i.e. one probability value for 1=female and one probability value for 0=male.
- * The size of gender, male and female, are n_1 and n_2 , respectively. As previously defined, for the two scenarios $N = 1200$, $N = 2400$: n_1 has 716 and 1290 and n_2 has 484 and 1110, respectively. In this simulation, the total number of observations $N = n_1 + n_2 = 1200$ and 2400, respectively.
- * Using the probability matrix and the sample size, a multinomial nominal response (voter intent) matrix with count y_{ij} is randomly generated through the **rmultinom** R function.

As part of the study with 1000 simulations, two particular simulation counts for voter intent to select a party are chosen for total sample sizes: $N = 1200$ and $N = 2400$, respectively.

Table 3.1: # of voter by gender for $N = 1200$

	Vote intent			
	CP	LP	Other	Total
Male	160	352	204	716
Female	109	55	320	484
Total	269	407	524	1200

Table 3.2: # of voter by gender for $N = 2400$

	Vote intent			
	CP	LP	Other	Total
Male	303	637	350	1290
Female	275	126	709	1110
Total	578	763	1059	2400

We simulate 1000 times for both scenarios, and use the NR coded in R to obtain the parameter estimate ($\hat{\beta}_{ij}$) which is the coefficient for the i^{th} subject and j^{th} category with i representing male and female and j representing the CP or LP.

From the simulation results, we calculate the descriptive statistics for the log-likelihood and parameter estimates $\hat{\beta}_{ij}$, and we calculate the mean square error (MSE), bias, estimated coverage probability (ECP) and the confidence interval average width (CIAW) for parameter estimates $\hat{\beta}_{ij}$ as seen in Tables 3.3 and 3.4. The green rows in the tables below show the true parameter values.

Table 3.3: Descriptive Statistics, Bias, MSE, ECP and CIAW for MLEs when $N = 1200$

Descriptive Statistics	logLik	Parameter estimates			
		β_{01}	β_{11}	β_{02}	β_{12}
Min	-1211.795	-0.4629	-1.3184	0.3340	-3.0394
Q_1	-1173.937	-0.2350	-0.9919	0.5337	-2.4478
Range	104.167	0.6188	0.9093	0.5772	1.2880
Median	-1162.732	-0.1649	-0.8821	0.5889	-2.3349
Mean	-1162.328	-0.1631	-0.8849	0.5925	-2.3404
St. Dev	17.081	0.1052	0.1510	0.0885	0.1735
IQR	22.822	0.1458	0.2132	0.1181	0.2289
Q_3	-1151.115	-0.0892	-0.7787	0.6518	-2.2189
Max	-1107.629	0.1559	-0.4090	0.9112	-1.7514
Bias	NA	-0.0061	0.0001	-0.0005	-0.0074
MSE	NA	0.0111	0.0228	0.0078	0.0302
True β 's	NA	-0.157	-0.885	0.593	-2.333
90% ECP	NA	90.30	91.10	90.50	89.60
95% ECP	NA	94.90	96.00	95.40	94.80
99% ECP	NA	99.20	99.10	99.00	98.70
90% CIAW	NA	0.3475	0.5027	0.2934	0.5636
95% CIAW	NA	0.4141	0.5990	0.3496	0.6716
99% CIAW	NA	0.5442	0.7872	0.4595	0.8826

The simulation results for $N = 1200$ show that all the parameter estimates are unbiased. The histograms of the MLEs as shown in Figure 3.2 demonstrate that the properties of MLEs (consistency and asymptotic normality) hold true and that our results satisfy these properties. The ECP for the CIs 90%, 95% and 99% are all close to their respective confidence levels as the ECPs are within approximately 1% of the nominal confidence levels, indicating that the coverage is quite accurate.

Table 3.4: Descriptive Statistics, Bias, MSE, ECP and CIAW for MLEs when $N = 2400$

Descriptive Statistics	logLik	Parameter estimates			
		β_{01}	β_{11}	β_{02}	β_{12}
Min	-2365.642	-0.3867	-1.2646	0.3662	-2.6770
Q_1	-2320.440	-0.2092	-0.9520	0.5468	-2.4046
Range	147.590	0.4766	0.7144	0.4407	0.6802
Median	-2303.852	-0.1594	-0.8761	0.5935	-2.3311
Mean	-2303.969	-0.1574	-0.8833	0.5921	-2.3293
St. Dev	23.749	0.0766	0.1052	0.0675	0.1141
IQR	31.523	0.1012	0.1355	0.0890	0.1572
Q_3	-2288.917	-0.1079	-0.8166	0.6358	-2.2474
Max	-2218.052	0.0899	-0.5502	0.8069	-1.9968
Bias	NA	-0.0004	0.0017	-0.0009	0.0037
MSE	NA	0.0059	0.0111	0.0046	0.0130
True β 's	NA	-0.157	-0.885	0.593	-2.333
90% ECP	NA	91.60	90.90	88.70	91.10
95% ECP	NA	95.50	95.10	94.00	95.80
99% ECP	NA	98.70	98.90	99.30	99.40
90% CIAW	NA	0.2585	0.3522	0.2186	0.3843
95% CIAW	NA	0.3080	0.4196	0.2605	0.4580
99% CIAW	NA	0.4047	0.5515	0.3423	0.6018

The simulation results for $N = 2400$ also show that all the parameter estimates are unbiased. The histograms of the MLEs as shown in Figure 3.2 demonstrate that the properties of MLEs (consistency and asymptotic normality) hold true and that our results satisfy these properties. The ECP for the CIs 90%, 95% and 99% are all close to their respective confidence levels as the ECPs are within approximately 1% of the nominal confidence levels, indicating that the coverage is quite accurate.

3.5.1 The Effect of Increase in Sample Size

We compare the two scenarios for $N = 1200$ and $N = 2400$ and note the following:

- * For the different sample sizes, both sets of simulations show that we have unbiased estimators. The bias for $N = 1200$ is slightly higher than the bias for $N = 2400$. As we would expect, as the sample size increases, the bias decreases.
- * When the sample size is increased, the MSE decreases. This behaviour shows that the procedure is working as it was intended to.
- * As expected, we have the same ECP regardless of sample size and we can conclude that our procedure is working well. This also demonstrates that there is no over or under estimation of the parameters.
- * The width of a CI is related to the sample size and its coverage probability - wider CIs have higher coverage probabilities, narrower CIs have lower coverage probabilities.

According to the MLE properties, all estimators are asymptotically normally distributed. All the Kernel density and histograms of estimates $\hat{\beta}_{ij}$ look normal, as seen in Figure 3.2:

Figure 3.2: Kernel density and histograms of unconstrained MLEs $\hat{\beta}_{ij}$

Chapter 4

Constrained Statistical Inference

4.1 Introduction

This section provides a brief review of constrained statistical inference and optimization techniques used throughout the thesis. There is often a need to use constrained statistical inference (CSI) in many areas, including research, science, technology, and others. Various types of data require the use of constraints in the space of the unknown parameters. Using constraints in statistical modeling allows for additional estimation and hypothesis tests and can make modeling techniques more difficult; however, by using them, we ensure the use of statistical information which may have otherwise been ignored. Incorporating prior information, in turn, can lead to improvements and efficiencies as well as improve statistical analysis. CSI has become more popular for this reason, not only in the statistical community, but also in multidisciplinary fields where the application of CSI has proven beneficial. For example, a hockey team owner may use CSI to determine whether or not choosing players with a high ranking in the entry draft will improve their team's performance [9]. Constrained statistical inference (also known as: one-sided testing, isotonic monotone regression, or restricted analysis) comes with advantages such as natural constraints imposed by the situation. Inference with constraints is often more efficient than unrestricted methods which ignore the constraints. Additionally, restricted maximum likelihood estimation is also efficient as it obtains consistent estimates of parameters in most cases.

Constraints on the parameter space(s) can come in the form of ordering or inequalities as well as observations and experimental outcomes (responses). Ordering can be partial, linear, non-linear, total or implicit. Inequalities can be linear or non-linear. While there are many

applications of CSI, this work focuses mainly on ordering and inequalities. The main idea is to incorporate the inequality and order restrictions on the statistical model parameters. When setting up hypothesis testing, we know that we will achieve more powerful results with a one-sided test than with regular two-sided testing in cases where there is a naturally-occurring constraint.

For the most part, data are normally distributed, which calls for linear modeling techniques; however, there are also many instances where this is not the case and the mean of an observation is not a linear combination of parameters, such as binary, Poisson and multinomial regression models. The methods used to compute the restricted maximum likelihood estimation (RMLE) are shown in Section 4.3.2. We will cover a globally convergent algorithm [11], based on gradient projections, for maximum likelihood estimation under linear equality and inequality constraints on parameters.

There is extensive literature on estimation and hypothesis testing for GLMs and well-developed statistical software. Many disciplines use logistic, loglinear and probit regression, which are all characterized by a nonlinear link function. This link function linearly relates the parameters of interest to a number of explanatory variables. For parameter estimation and associated hypothesis testing, we use the link function in various inference methods, such as maximum likelihood, quasi-likelihood and generalized estimating equations (GEE). Although estimating parameters using MLE is preferable, it can be complex and computationally intensive; and while we can use other methods (such as integral approximations (quasi-likelihood), stochastic methods such as Monte Carlo methods, and GEE) that are less computationally intensive, these various methods have shown inconsistencies that outweigh their computational benefits. For this reason, this thesis will explore maximum likelihood inference techniques.

To expand on the existing literature regarding unrestricted inference for GLMs, this thesis aims to extend from binary GLM and GLMM to a multinomial logit, multivariate GLM and multivariate GLMM, subject to ordered equality and inequality constraints. We extend the MLE and likelihood ratio hypothesis testing methods for the binary and MGLM and MGLMM

subject to linear equality and inequality constraints on the parameters of interest. Dr. Karelyn Davis' work [10] covers restricted estimation under binary and Poisson data. As shown in Section 6.2.1, Dr. Karelyn Davis' results were replicated using complete model matrices (without missing data).

4.2 Concepts and Definitions

This section defines important technical terms which will be used throughout the paper. These are directly referenced from [10] and [15]. Let $\mathbb{R}^p$ denote the p -dimensional Euclidean space.

Definition 4.1 (Convex Set): A set $\mathcal{A} \subset \mathbb{R}^p$ is said to be convex if $\{\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\} \in \mathcal{A}$ whenever $\mathbf{x}, \mathbf{y} \in \mathcal{A}$ and $0 < \alpha < 1$. Therefore, $\mathcal{A}$ is a convex set if the line segment joining $\mathbf{x}$ and $\mathbf{y}$ is in $\mathcal{A}$ whenever the points $\mathbf{x}$ and $\mathbf{y}$ are in $\mathcal{A}$. We generalize this to more than two vectors by saying $\mathbf{z}$ is a convex combination of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ if there are some $\alpha_1, \alpha_2, \dots, \alpha_p$ such that $\mathbf{z} = \alpha_1\mathbf{x}_1, \alpha_2\mathbf{x}_2, \dots, \alpha_p\mathbf{x}_p$ and $\sum_{\ell}^p \alpha_{\ell} = 1$.

Definition 4.2 (Cone with vertex): A set $\mathcal{A}$ is said to be a *cone with vertex* $\mathbf{x}_0$ if $\mathbf{x}_0 + k(\mathbf{x} - \mathbf{x}_0) \in \mathcal{A}$ for every $\mathbf{x} \in \mathcal{A}$ and $k \geq 0$. If the vertex is the origin, then we refer to the set $\mathcal{A}$ as a *cone*. A cone is a set that consists of infinite straight lines starting from the origin.

Definition 4.3 (Inner product on $\mathbb{R}^p$): Let $\mathbf{V}$ be a $p \times p$ symmetric positive definite matrix, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Then $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}} = \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y}$ defines an inner product on $\mathbb{R}^p$. From this, the (*Euclidean*) *norm* of vector $\mathbf{x} \in \mathbb{R}^p$ is $\|\mathbf{x}\|_{\mathbf{V}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{V}}}$ and the corresponding distance between $\mathbf{x}$ and $\mathbf{y}$ is $\|\mathbf{x} - \mathbf{y}\|_{\mathbf{V}}$.

If $\mathbf{x}^T \mathbf{V}^{-1} \mathbf{y} = 0$ then we say that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal with respect to (w.r.t) $\mathbf{V}$. "Orthogonal w.r.t $\mathbf{V}$ " may be abbreviated by $\mathbf{V}$ -orthogonal. If $\mathbf{x}$ and $\mathbf{y}$ are orthogonal w.r.t $\mathbf{V}$ then a version of the Pythagoras theorem holds: $\|\mathbf{x} + \mathbf{y}\|_{\mathbf{V}}^2 = \|\mathbf{x}\|_{\mathbf{V}}^2 + \|\mathbf{y}\|_{\mathbf{V}}^2$. A consequence of this is that the shortest distance between a point and a plane is the distance from the point to the place along a line is orthogonal to the plane.

Definition 4.4 (Feasible Regions and Polyhedron): Let the matrix $\mathbf{A} \in \mathbb{R}^{m \times p}$ and vector $\mathbf{b} \in \mathbb{R}^m$ specify a set of m inequality constraints, one for each row of $\mathbf{A}$. The i^{th} constraint comes from the i^{th} row and is $\mathbf{a}_i^T \boldsymbol{\beta} \leq b_i$.

- (1) The set of vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ that satisfies $\mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}$ is called the feasible region; and
- (2) The set of vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ that satisfies inequality and equality *constraints* of the form $\mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}$ and $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ is called a *polyhedron*.
- (3) **Polyhedron** [17] is a solution set of finitely many linear inequality and equality.
- (4) Let $\mathbf{a}_1, \dots, \mathbf{a}_m$ be m points in $\mathbb{R}^p$ and $\mathcal{P} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \mathbf{a}_i^T \boldsymbol{\beta} \geq 0 \text{ for } i = 1, \dots, m\}$. Then $\mathcal{P}$ is a closed convex cone and is called a *polyhedral cone*. Note that $\mathcal{P}$ is the intersection of a finite number of the half-spaces and hyper-planes, $\{\boldsymbol{\beta} : \mathbf{a}_1^T \boldsymbol{\beta} \geq 0\}, \dots$, and $\{\boldsymbol{\beta} : \mathbf{a}_m^T \boldsymbol{\beta} \geq 0\}$.

Figure 4.1: Polyhedron $\mathcal{P}$ (shown shaded) is the intersection of five half-spaces, with outward normal vectors $a_1, \dots, a_5$.

In three dimensions, the boundaries of the set are formed by “flat faces”, whereas in two dimensions, the boundaries are formed by line segments, and a polyhedron is a polygon. Since the feasible region of a linear program is a polyhedron, a linear program involves optimization of a linear objective function over a polyhedral feasible region. We can consider a polyhedron as the intersection of a collection of half-spaces, where a half-space is a set of points that

satisfy a single inequality constraint. So, each constraint $\mathbf{a}_i^T \boldsymbol{\beta} \leq b_i$ defines a half-space, and the polyhedron characterized by $\mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}$ is the intersection of m such half-spaces.

Definition 4.5: Let C be a closed convex set in $\mathbb{R}^p$ and $\mathbf{x} \in \mathbb{R}^p$. Let $\tilde{\boldsymbol{\theta}}$ be the point in C that is closest to $\mathbf{x}$ w.r.t the distance $\|\cdot\|_{\mathbf{V}}$, i.e.

$$\|\mathbf{x} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}} = \min_{\boldsymbol{\theta} \in C} (\mathbf{x} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\theta}).$$

The vector $\tilde{\boldsymbol{\theta}} = P_{\mathbf{V}}(\mathbf{x}|C)$ is the projection of $\mathbf{x}$ onto C , thus

$$\tilde{\boldsymbol{\theta}} = P_{\mathbf{V}}(\mathbf{x}|C) = \min_{\boldsymbol{\theta} \in C} (\mathbf{x} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\theta}).$$

Definition 4.6 (Dual and Polar cone): Let C be a cone, the **dual cone** C^* of C is the set defined as

$$C^* = \{\mathbf{y} \in \mathbb{R}^p : \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y} \geq 0 \quad \forall \mathbf{x} \in C\}, \text{ w.r.t the inner product } \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y}.$$

Clearly,

- * C^* is always a closed set, regardless if C is closed or not,
- * C^* is also always a convex cone regardless if C is convex or not, and
- * geometrically, $\mathbf{y} \in C^*$ iff $-\mathbf{y}$ is the normal of a hyperplane that supports C at the origin.

The **polar cone** of C is the set defined as

$$C^o = \{\mathbf{y} \in \mathbb{R}^p : \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y} \leq 0 \quad \forall \mathbf{x} \in C\}, \text{ w.r.t the inner product } \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y}.$$

Clearly:

- * C^o is the collection of vectors which do not form an acute angle with any vector in C .
- * The polar cone is equal to the negative of the dual cone, i.e. $C^o = -C^*$.
- * The boundaries of C^o are perpendiculars to the boundaries of C .

4.3 Constrained Optimization

Consider maximizing the objective function $f(\boldsymbol{\theta})$ subject to the equality and inequality constraints:

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \quad f(\boldsymbol{\theta}) \quad (4.1)$$

$$\text{subject to} \quad h_i^*(\boldsymbol{\theta}) = b_i, \quad i = 1, \dots, n_e, \quad (4.2)$$

$$g_j^*(\boldsymbol{\theta}) \leq b_j, \quad j = 1, \dots, n_g. \quad (4.3)$$

The above expansion can also be written as

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \quad f(\boldsymbol{\theta}) \quad (4.4)$$

$$\text{subject to} \quad h_i(\boldsymbol{\theta}) = h_i^*(\boldsymbol{\theta}) - b_i = 0, \quad i = 1, \dots, n_e, \quad (4.5)$$

$$g_j(\boldsymbol{\theta}) = g_j^*(\boldsymbol{\theta}) - b_j \leq 0, \quad j = 1, \dots, n_g, \quad (4.6)$$

where

- * $\boldsymbol{\theta}$ is the parameter/decision/optimization variable,
- * $f(\boldsymbol{\theta})$ is the utility or objective function,
- * $h_i(\boldsymbol{\theta})$ are the equality constraint functions, $i = 1, \dots, n_e$, and
- * $g_j(\boldsymbol{\theta})$ are the inequality constraint functions, $j = 1, \dots, n_g$.
- * The functions $f(\boldsymbol{\theta}), h_i(\boldsymbol{\theta}), g_j(\boldsymbol{\theta})$ are differentiable concave functions.

Let $\tilde{\boldsymbol{\theta}}$ be the optimal solution.

- * If an inequality constraint holds with equality at $\tilde{\boldsymbol{\theta}}$, i.e. $g_j(\tilde{\boldsymbol{\theta}}) = 0$, we call the constraint **binding or active**, which means there is no movement in the direction of the constraint.

Based on our definition of a polyhedron in section 4.2, when a constraint is active, the point in question lies on the hyperplane forming the border of the half space. In three dimensions, that point must lie on the surface associated with the constraint, and in two dimensions, it must lie on the edge (or line segment) associated with the constraint.

- * If an inequality constraint holds as a strict inequality at $\tilde{\theta}$, i.e. $g_j(\tilde{\theta}) < 0$, we call the constraint **non-binding or inactive**, which means the point varies in the direction of the constraint. If a constraint is non-binding, the optimization problem could have the same solution even in the absence of that constraint.

In the design space, we have two domains:

- * the feasible domain where the constraints are satisfied, and
- * the infeasible domain where at least one of the constraints is violated.
- * The feasible domain is convex if all the inequality constraints g_j are concave (that is, $-g_j$ are convex) and the equality constraints are linear.

In most cases, we can find the maximum between these two domains, where the inequality constraint holds equality: where $g_j(\theta) = 0$ for at least one j . If not, the inequality constraints may be removed without making changes to the solution.

Theorem 4.1: *Given a polyhedron $\{\theta \in \mathbb{R}^p \mid \mathbf{A}\theta \leq \mathbf{b}\}$ for some $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{b} \in \mathbb{R}^m$, any set of p linearly independent active constraints identifies a unique basic solution.*

To understand the optimization method for equality and inequality constraints, we will briefly review the Kuhn-Tucker (KT) conditions, also known as Karush-Kuhn-Tucker conditions (KKT conditions), and then consider the Gradient Projection Theory.

4.3.1 Kuhn-Tucker(KT) Conditions

In general, any given maximizing problem (4.4) may have several local maxima. Only under special circumstances can we be sure of the existence of a single global minimum. The necessary

conditions for a maximum of the constrained problem are obtained by using the Lagrange multiplier method. We define the **Lagrangian function**:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\boldsymbol{\theta}) + \sum_{i=1}^{n_e} \lambda_i h_i(\boldsymbol{\theta}) + \sum_{j=1}^{n_g} \nu_j g_j(\boldsymbol{\theta}) \quad (4.7)$$

In matrix form:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\boldsymbol{\theta}) + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\theta}) + \boldsymbol{\nu}^T \mathbf{g}(\boldsymbol{\theta}), \quad (4.8)$$

where

- * $h_i(\boldsymbol{\theta})$ are differentiable concave functions and are components of the $n_e \times 1$ vector-valued functions $\mathbf{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), h_2(\boldsymbol{\theta}), \dots, h_{n_e}(\boldsymbol{\theta}))$,
- * λ_i are Lagrange multipliers and are components of the $n_e \times 1$ vector $\boldsymbol{\lambda}$,
- * $g_j(\boldsymbol{\theta})$ are differentiable concave functions and are components of the $n_g \times 1$ vector-valued functions $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \dots, g_{n_g}(\boldsymbol{\theta}))$, and
- * ν_j are Lagrange multipliers and are components of the $n_g \times 1$ vector $\boldsymbol{\nu}$.
- * The constraints $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}, \mathbf{g}(\boldsymbol{\theta}) \leq \mathbf{0}$ are referred to as *functional constraints*.

To test a point to see if it is a critical point in a constraints nonlinear, we use the KT conditions (see below). While these conditions will not determine if the point is the local maximum, minimum or the saddle point, they will determine if the point is a critical point or not. Note however, since the goal is to determine maximum or minimum, finding a point that satisfies the KT conditions is finding the local optimum. For local optimality we consider the following scenarios:

- (1) Unconstrained - when no constraints are active at the local optimum point, the gradient of the objective function will be zero at the optimal point. This is when the maximum

(hill) or minimum (valley) of the objective function is not near the limiting value of any constraint, and the Hessian can be used.

- * The gradient of f at $\tilde{\boldsymbol{\theta}}$ is zero: $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \mathbf{0}$.
- * At maximum, Hessian of f at $\tilde{\boldsymbol{\theta}}$ is negative semi-definite: $\mathbf{u}^T \nabla_{\boldsymbol{\theta}}^2 f(\boldsymbol{\theta}) \mathbf{u} \leq 0 \quad \forall \mathbf{u} \in \mathbb{R}^p$.
- * At minimum, Hessian of f at $\tilde{\boldsymbol{\theta}}$ is positive semi-definite: $\mathbf{u}^T \nabla_{\boldsymbol{\theta}}^2 f(\boldsymbol{\theta}) \mathbf{u} \geq 0 \quad \forall \mathbf{u} \in \mathbb{R}^p$.

- (2) Equality and Inequality constraints - when at least one constraint *is active* at the local optimum point, a constraint is preventing the improvement of the value of the objective function, the gradient is not zero ($\nabla f(\boldsymbol{\theta}) \neq \mathbf{0}$), and the Hessian cannot be used. The KT equivalence theorem will determine a $\tilde{\boldsymbol{\theta}}$ that maximizes/minimizes $f(\boldsymbol{\theta})$ constrained by $\mathbf{h}(\boldsymbol{\theta})$ and $\mathbf{g}(\boldsymbol{\theta})$. The KT conditions are outlined below:

- (1) Stationarity (i.e. No feasible descent) - ensures that no other feasible direction exists to potentially improve the objective function.

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\nu})}{\partial \boldsymbol{\theta}} = \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_{i=1}^{n_e} \lambda_i \frac{\partial h_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_{j=1}^{n_g} \nu_j \frac{\partial g_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

In Matrix notation, we write

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \boldsymbol{\lambda}^T \nabla_{\boldsymbol{\theta}} \mathbf{h}(\boldsymbol{\theta}) + \boldsymbol{\nu}^T \nabla_{\boldsymbol{\theta}} \mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}.$$

NOTE: We cannot use the $\partial f(\boldsymbol{\theta}) = \nabla f(\boldsymbol{\theta})$ for a differentiable function f unless f is convex.

- (2) Complementary slackness - applies only to inequality constraints. We apply a Lagrange multiplier in these cases: positive Lagrange multiplier if the constraint is active and set the Lagrange multiplier to zero if the constraint is inactive:

$$\nu_j g_j(\boldsymbol{\theta}) = 0 \text{ for all } j.$$

In Matrix notation, we write

$$\boldsymbol{\nu} \mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}.$$

- (3) Feasible constraints (i.e. primal feasibility) - applies to equality and inequality constraints. A vector $\boldsymbol{\theta}$ is feasible if it satisfies all given constraints, that is,

$$h_i(\boldsymbol{\theta}) = 0, \quad g_j(\boldsymbol{\theta}) \leq 0 \text{ for all } i, j.$$

In Matrix notation, we write

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}, \quad \mathbf{g}(\boldsymbol{\theta}) \leq \mathbf{0}.$$

- (4) Dual feasibility (i.e. Positive Lagrange Multipliers) - applies to inequality constraints, that is,

$$\nu_j \geq 0 \text{ for all } j.$$

In Matrix notation, we write

$$\boldsymbol{\nu} \geq \mathbf{0}.$$

Kuhn-Tucker provides us with the conditions needed for optimization; however, we must go further to identify the method or algorithm required to optimize our estimates. Finding the maximum using the Kuhn-Tucker conditions is difficult in many cases because we would need to consider many active and inactive constraint combinations, which would require solving for highly nonlinear equations. The gradient projection theory is used to handle these problems.

4.3.2 Gradient Projection Theory

Solving for restricted maximization problems can be done using many approaches; one popular approach is to use the Gradient Projection (GP) method. The GP method is inspired by the

ordinary method of generalized steepest ascent algorithm for unconstrained problems. To define the direction of movement $\mathbf{d}$, the gradient, or score function, of log-likelihood *is projected onto the working surface of active constraints*. Using GP, we will look at a globally convergent algorithm for the maximum likelihood estimation (MLE) under linear equality and inequality constraints on parameters. The MLE under parameter constraints is relevant to many statistical computing problems, which is why a globally convergent algorithm is of importance to statisticians. As Jamshidian explains, an algorithm is considered convergent if it converges to a local maximizer of a nonlinear functional from almost any starting value [11]. As stated, a constraint is considered active if it holds with equality. The GP algorithm covers linear and nonlinear constraints; this thesis will only focus on linear constraints.

Consider the restricted ML problem with the equality and inequality constraints:

$$\underset{\boldsymbol{\beta} \in \Omega}{\text{maximize}} \quad \ell(\boldsymbol{\beta}) \quad (4.9)$$

$$\text{subject to} \quad \mathbf{a}_i^T \boldsymbol{\beta} = b_i, \quad i \in I_1 \quad (4.10)$$

$$\mathbf{a}_i^T \boldsymbol{\beta} \leq b_i, \quad i \in I_2, \quad (4.11)$$

or in matrix form:

$$\underset{\boldsymbol{\beta} \in \Omega}{\text{maximize}} \quad \ell(\boldsymbol{\beta}) \quad (4.12)$$

$$\text{subject to} \quad \mathbf{A}_1 \boldsymbol{\beta} = \mathbf{b}_1 \quad (4.13)$$

$$\mathbf{A}_2 \boldsymbol{\beta} \leq \mathbf{b}_2, \quad (4.14)$$

where

- * $\ell(\boldsymbol{\beta})$ is the sufficiently smooth log-likelihood (objective) function,
- * $\boldsymbol{\beta}$ is a $p \times 1$ vector of true parameters,
- * $\Omega = \Omega_1 \cup \Omega_2$ is the constrained parameter space, where $\Omega_1 = \{\boldsymbol{\beta} \in \mathbb{R}^p : \mathbf{A}_1 \boldsymbol{\beta} = \mathbf{b}_1\}$ representing equality constraints, and where $\Omega_2 = \{\boldsymbol{\beta} \in \mathbb{R}^p : \mathbf{A}_2 \boldsymbol{\beta} \leq \mathbf{b}_2\}$ representing

inequality constraints,

- * $\mathbf{A}_1$ and $\mathbf{A}_2$ are $m_1 \times p$ and $m_2 \times p$ known matrices of full rank ($m_1 \leq p$) and ($m_2 \leq p$), whose rows consist of constraints $\mathbf{a}_i^T$ for all $i \in I_1$ and $i \in I_2$, respectively,
- * $\mathbf{b}_1$ and $\mathbf{b}_2$ are $m_1 \times 1$ and $m_2 \times 1$ known right-side vectors of constraints whose component b_i is a given scalar for all $i \in I_1$ and $i \in I_2$, respectively,
- * I_1 is an index set of equality constraints, and
- * I_2 is an index set of inequality constraints.

For maximization problems, we need to derive the gradient (score function) for the log-likelihood as $S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Also the generalized gradient $\tilde{S}(\boldsymbol{\beta}) = \mathbf{W}^{-1}S(\boldsymbol{\beta})$ of $\ell(\boldsymbol{\beta})$ is derived in the metric (distance) of a positive definite symmetric matrix $\mathbf{W}$, which is also known as the Gram matrix, Gramian matrix or Gramian. $\mathbf{W}$ is defined by the p -dimensional Euclidean space $\mathbb{E} = \mathbb{R}^p$ with the inner product of two vectors:

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle_W = \mathbf{v}_1^T \mathbf{W} \mathbf{v}_2. \quad (4.15)$$

Suppose we have m number of active constraints satisfying $\mathbf{a}_i^T \boldsymbol{\beta} = b_i$ and some inactive constraints $\mathbf{a}_i^T \boldsymbol{\beta} < b_i$ at a given feasible point $\boldsymbol{\beta}$. To maximize the log-likelihood function $\ell(\boldsymbol{\beta})$ subject to (4.10) and (4.11), we run the GP algorithm. Start with the initial working set $\mathbb{W}$.

- * $\mathbb{W}$ is the working set of active constraints. It contains the constraint indexes of I_1 , if there are any, and may contain any constraint indexes from I_2 as well.
- * Let $\mathbf{A}$ be an $m \times p$ matrix with rows $\mathbf{a}_i^T$ for all $i \in \mathbb{W}$ of rank $m < p$. The matrix $\mathbf{A}$ is composed of the rows of working constraints, i.e. contain matrix $\mathbf{A}_1$, if any, and some rows of $\mathbf{A}_2$ if they are active.
- * Let $\mathbf{b}$ be the corresponding right-side vector of b_i 's.

* Let $\boldsymbol{\lambda} \in \mathbb{R}^m$ be the Lagrangian multiplier.

If the set of active constraints at the optimal point is determined, then the problem is essentially an equality constraint one. *Active set method* is a procedure that determines optimal active constraints by moving among several working sets of potential optimal active constraints [12]. We define the **Lagrangian function** for maximization as:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \ell(\boldsymbol{\beta}) + \boldsymbol{\lambda}^T (\mathbf{A}\boldsymbol{\beta} - \mathbf{b}). \quad (4.16)$$

Based on the Global Convergence Theorem [13], the GP algorithm is considered globally convergent since it is a generalized steepest ascent algorithm. Also, since $\mathbf{d}$ is a feasible and ascent direction, even a small step from $\boldsymbol{\beta}_r \in \Omega$ in the direction of $\mathbf{d}$ will give a new feasible point $\tilde{\boldsymbol{\beta}}_r = \boldsymbol{\beta}_r + \mathbf{d}$ such that $\ell(\tilde{\boldsymbol{\beta}}_r) > \ell(\boldsymbol{\beta}_r)$. Note, we use subscript r to indicate the point is in the feasible, or restricted, space.

The movement in direction $\mathbf{d}$ causes an increase in the log-likelihood $\ell(\boldsymbol{\beta})$, so the new feasible point $\tilde{\boldsymbol{\beta}}_r$ is determined by a feasible direction vector $\mathbf{d}$ that satisfies $\nabla \ell(\boldsymbol{\beta})\mathbf{d} = S(\boldsymbol{\beta})\mathbf{d} > 0$. The new point $\tilde{\boldsymbol{\beta}}_r$ is in Ω if and only if, $\mathbf{d}$ is in the null space of $\mathbf{A}$. Consider the space of feasible directions: $\mathcal{N} = \{\mathbf{d} \in \mathbb{E} : \mathbf{A}\mathbf{d} = \mathbf{0}\}$. The set $\mathcal{N}$ defined by the working set of constraints is called the tangent subspace, which plays a role similar to the Hessian matrix in an unconstrained case. For all the working constraints to remain active, the directions must satisfy: $\mathbf{a}_i^T \mathbf{d} = 0, i \in \mathbb{W}$. To find the feasible solution satisfying the active constraints, the iterations in the GP algorithm start at a point in Ω to generate a set of feasible points by moving along $\mathbf{d}$, where $\mathbf{d}$ is obtained at a point $\boldsymbol{\beta}_r \in \Omega$ by projecting $\tilde{S}(\boldsymbol{\beta})$ onto $\mathcal{N}$ in the metric of $\mathbf{W}$.

When the direction vectors $\mathbf{d}$ lie in the tangent subspace $\mathcal{N}$, then we can consider a space composed of row vectors (active constraints) of matrix $\mathbf{A}$, which is defined as $\mathcal{O} = \{\mathbf{u} \in \mathbb{E} : \mathbf{u} = \mathbf{W}^{-1}\mathbf{A}^T\boldsymbol{\lambda} \text{ for some } \boldsymbol{\lambda} \in \mathbb{R}^m\}$. $\mathcal{O}$ is orthogonal and conjugate to the tangent space $\mathcal{N}$ in the metric of $\mathbf{W}$. Since $\mathcal{O}$ and $\mathcal{N}$ are conjugate, any vector can be written as the sum of vectors from each of these two complementary subspaces. Specifically, the generalized gradient vector,

where $\mathbf{d} \in \mathcal{N}$ and $\boldsymbol{\lambda} \in \mathbb{R}^m$ can be written as:

$$\tilde{S}(\boldsymbol{\beta}) = \mathbf{d} + \mathbf{W}^{-1} \mathbf{A}^T \boldsymbol{\lambda}. \quad (4.17)$$

Using the requirement that $\mathbf{A}\mathbf{d} = \mathbf{0}$ and the $\text{rank}(\mathbf{A}) = 0$. Multiplying both sides of (4.17) by the constraints matrix $\mathbf{A}$ we can solve for $\boldsymbol{\lambda}$. Thus

$$\mathbf{A}\mathbf{d} = \mathbf{A}\tilde{S}(\boldsymbol{\beta}) - \mathbf{A}\mathbf{W}^{-1} \mathbf{A}^T \boldsymbol{\lambda} = \mathbf{0} \Rightarrow \mathbf{A}\tilde{S}(\boldsymbol{\beta}) = (\mathbf{A}\mathbf{W}^{-1} \mathbf{A}^T) \boldsymbol{\lambda},$$

which leads to

$$\boldsymbol{\lambda} = (\mathbf{A}\mathbf{W}^{-1} \mathbf{A}^T)^{-1} \mathbf{A}\tilde{S}(\boldsymbol{\beta}), \quad (4.18)$$

and substitute $\boldsymbol{\lambda}$ from the equation (4.18) in gradient vector equation (4.17)

$$\tilde{S}(\boldsymbol{\beta}) = \mathbf{d} + \mathbf{W}^{-1} \mathbf{A}^T (\mathbf{A}\mathbf{W}^{-1} \mathbf{A}^T)^{-1} \mathbf{A}\tilde{S}(\boldsymbol{\beta}).$$

We can solve for the direction vector $\mathbf{d}$ as

$$\mathbf{d} = \mathbf{P}_{\mathbf{w}} \tilde{S}(\boldsymbol{\beta}), \text{ with} \quad (4.19)$$

$$\mathbf{P}_{\mathbf{w}} = \mathbf{I} - \mathbf{W}^{-1} \mathbf{A}^T (\mathbf{A}\mathbf{W}^{-1} \mathbf{A}^T)^{-1} \mathbf{A}, \quad (4.20)$$

where

* $\mathbf{I}$ is an $p \times p$ identity matrix,

* $\langle \tilde{S} - \mathbf{d}, \mathbf{d} \rangle_{\mathbf{w}} = 0$ i.e. $\tilde{S} - \mathbf{d}$ is orthogonal to $\mathbf{d}$, so we obtain

$$S^T \mathbf{d} = (S^T + \mathbf{d}^T - \mathbf{d}^T) \mathbf{d} = \|\mathbf{d}\|^2,$$

- * the direction $\mathbf{d}$ is the projection of $\tilde{S}$ onto $\mathcal{N}$ in the metric of $\mathbf{W}$,
- * $\mathbf{P}_w$ is a $p \times p$ projection matrix onto $\mathcal{N}$ in the metric $\mathbf{W}$ defined by the inner product (4.15), and
- * $\mathbf{P}_w$ is idempotent and self-adjoint.

Theorem 4.2 (Jamshidian, 2004): Consider the direction $\mathbf{d}$ defined in equation (4.19).

Then

- (1) If $\mathbf{d} \neq \mathbf{0}$, then $\mathbf{d}$ is an ascent and feasible direction at β_r w.r.t. the log-likelihood function $\ell(\cdot)$, and
- (2) the direction $\mathbf{d}$ is a generalized gradient of the log-likelihood function $\ell(\cdot)$ in $\mathcal{N}$ in the metric of $\mathbf{W}$ defined by the inner product (4.15).

If the projected gradient $\mathbf{d} = \mathbf{0}$ at a point β_r , then by (4.17) we have

$$W^{-1} [\mathbf{A}^T \boldsymbol{\lambda} - S(\beta_r)] = \mathbf{0} \Rightarrow S(\beta_r) = \mathbf{A}^T \boldsymbol{\lambda},$$

and the point β_r satisfies the necessary conditions for a maximum of ℓ in Ω , on the working surface. If the components of $\boldsymbol{\lambda}$ computed by (4.18) corresponding to the active inequalities $\mathbf{a}_i^T \boldsymbol{\beta} \leq b_i$ are all non-negative, then this coupled with $\mathbf{A}^T \boldsymbol{\lambda} - S(\beta_r) = \mathbf{0}$ means that the KT conditions for the original problem are satisfied at β_r and the GP algorithm terminates. If at least one of the components of $\boldsymbol{\lambda}$ is negative, we can relax the corresponding inequality and move to a new improved feasible point in a new direction obtained by projecting the gradient onto the subspace of remaining $m - 1$ active constraints [13].

To ensure that the log-likelihood is moving toward the maximum, we consider the selection of step size α_2 . As α_2 increases from zero, the new point computed as $\tilde{\beta}_r = \beta_r + \alpha_2 \mathbf{d}$, remains feasible at the onset and the log-likelihood $\ell(\beta_r)$ at that point will increase. At the point β_r , find the length of the feasible line segment and maximize the log-likelihood $\ell(\beta_r)$ corresponding

to this segment. If the maximum occurs at the boundary to the working set $\mathbb{W}$, a new constraint will become active and will be added to the working set.

The steps below summarize the GP-active set algorithm: Given an initial feasible point $\boldsymbol{\beta}_r \in \Omega$, which means $\mathbf{A}\boldsymbol{\beta}_r = \mathbf{b}$, the GP algorithm iterates through the steps below until convergence to $\hat{\boldsymbol{\beta}}$:

- Step 1)** Define the subspace of active constraints $\mathcal{N}$, and create the constraint matrix $\mathbf{A}$, and the working set $\mathbb{W}$.
- Step 2)** Calculate the projection matrix $\mathbf{P}_w = \mathbf{I} - \mathbf{W}^{-1}\mathbf{A}^T (\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T)^{-1} \mathbf{A}$ onto $\mathcal{N}$, and direction vector $\mathbf{d} = \mathbf{P}_w \tilde{S}(\boldsymbol{\beta}) = \mathbf{P}_w \mathbf{W}^{-1} \nabla \ell(\boldsymbol{\beta})$.
- Step 3)** If $\mathbf{d} = \mathbf{0}$, find the Lagrange multipliers $\boldsymbol{\lambda} = (\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T)^{-1} \mathbf{A} \tilde{S}(\boldsymbol{\beta})$ with component λ_i , where i is the row index of the constraint matrix $\mathbf{A}$.
 - a)** If all the components of $\boldsymbol{\lambda}$ are positive, i.e. $\lambda_i \geq 0$ for $i \in \mathbb{W} \cap I_2$, associated to the active inequalities, **stop**; and declare that the KT necessary conditions are satisfied at the point $\boldsymbol{\beta}_r$.
 - b)** If at least one component of $\boldsymbol{\lambda}$ for $i \in \mathbb{W} \cap I_2$ is negative, find the index of the smallest negative component of $\boldsymbol{\lambda}$ and remove it from the set $\mathbb{W}$. Drop a row from both $\mathbf{A}$, and $\mathbf{b}$, and return to **Step 2**.
- Step 4)** If $\mathbf{d} \neq \mathbf{0}$, search for α_1 and α_2 such that

$$\alpha_1 = \max_{\alpha} \{ \alpha : \boldsymbol{\beta} + \alpha \mathbf{d} \text{ is feasible} \} \text{ and}$$

$$\alpha_2 = \max_{\alpha} \{ \ell(\boldsymbol{\beta} + \alpha \mathbf{d}) : 0 \leq \alpha \leq \alpha_1 \}.$$

Set $\tilde{\boldsymbol{\beta}}_r = \boldsymbol{\beta}_r + \alpha_2 \mathbf{d}$ and return to **Step 1**, which means that we add new constraints, if any, to $\mathbf{A}$ and $\mathbf{b}$ that is on the boundary to the working set $\mathbb{W}$. Then update $\boldsymbol{\beta}_r$ using $\tilde{\boldsymbol{\beta}}_r$ and return to **Step 2**.

If the restricted ML problem (4.9) consists only of **equality** constraints, given an initial feasible point $\beta_r \in \Omega_1$, we can simplify the GP algorithm as follows:

- Step 1)** Compute $\mathbf{P}_w = \mathbf{I} - \mathbf{W}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T)^{-1}\mathbf{A}$ and $\mathbf{d} = \mathbf{P}_w\tilde{S}(\beta) = \mathbf{P}_w\mathbf{W}^{-1}\nabla\ell(\beta)$. If $\mathbf{d} = \mathbf{0}$, **stop**; and declare convergence.
- Step 2)** Choosing $\alpha = \max_{\alpha}\{\ell(\beta_r + \alpha\mathbf{d})\}$, set the new point $\tilde{\beta}_r = \beta_r + \alpha\mathbf{d}$. Instead, use step-halving as a substitute to α , such that $\ell(\tilde{\beta}) > \ell(\beta)$ for $\tilde{\beta}_r = \beta_r + (0.5)^k\mathbf{d}$, to obtain the smallest integer k , where $k \geq 0$.
- Step 3)** Update β_r using $\tilde{\beta}_r$ and return to **Step 1**.

We present below some useful definitions.

Definition 4.7 (Feasible direction): A vector $\mathbf{d} \in \mathbb{R}^n, \mathbf{d} \neq \mathbf{0}$ is said to be a *feasible direction* at $\mathbf{x} \in \mathbf{X}$ if there exists $\delta_1 > 0$ such that $\mathbf{x} + \alpha\mathbf{d} \in \mathbf{X}$ for all $\alpha \in (0, \delta_1)$. Let $\mathcal{F}(\mathbf{x})$ be the set of feasible directions at $\mathbf{x} \in \mathbf{X}$.

Definition 4.8 (Descent direction): A vector $\mathbf{d} \in \mathbb{R}^n, \mathbf{d} \neq \mathbf{0}$ is said to be a *descent direction* at $\mathbf{x} \in \mathbf{X}$ if there exists $\delta_2 > 0$ such that $f(\mathbf{x} + \alpha\mathbf{d}) < f(\mathbf{x})$ for all $\alpha \in (0, \delta_2)$. Let $\mathcal{D}(\mathbf{x})$ be the set of descending directions at $\mathbf{x} \in \mathbf{X}$.

Theorem 4.3: Let $\mathbf{X}$ be a non-empty set in $\mathbb{R}^n$ and $\mathbf{x}^* \in \mathbf{X}$ be the local optimal of f over $\mathbf{X}$. Then, $\mathcal{F}(\mathbf{x}^*) \cap \mathcal{D}(\mathbf{x}^*) = \phi$

Definition 4.9 (Regular point): If the gradient vectors of the active constraints are linearly independent at point β_r , which satisfies the equality constraint, then the point β_r is called a regular point.

In the next chapter, we describe the GP algorithm in the context of the multivariate normal distribution.

Chapter 5

Inference for Multivariate Normal under Linear Inequality Constraints

5.1 Order Restricted/Constrained Inference

Pioneer and development work on constrained statistical inference began as early as the 1950s, culminating with the work done by Barlow, Bartholomew, Bremner and Brunk (1972). From there, researchers expanded and further developed the concepts and methodologies gradually, with works by Robertson, Wright and Dykstra (1988). Since then, many have followed suit, each adding their own contributions to the field.

Order restricted statistical inference is a statistical technique that deals with estimation or testing problems under equality and inequality constraints such as: order restriction, monotone function, and stochastic ordering.

- Order restriction: $\theta_1 \leq \theta_2 \leq \dots \leq \theta_k$
- Monotone function:

$$x_1 \leq x_2 \leq \dots \leq x_k \Rightarrow f(x_1) \leq f(x_2) \leq \dots \leq f(x_k)$$

- Stochastic ordering:

Let F and G be the cumulative density function (cdf) then $F(x) \leq G(x) \quad \forall x$

Define

$$C_0 = \{\mathbf{z} \in \mathbb{R}^k : z_1 = z_2 = \cdots = z_k\}$$

$$C_1 = \{\mathbf{z} \in \mathbb{R}^k : z_1 \leq z_2 \leq \cdots \leq z_k\}$$

$$C_2 = \{\mathbf{z} \in \mathbb{R}^k : \text{no restriction on } \mathbf{z}\}$$

Example 5.1 (One-Way ANOVA): The one-way ANOVA model is defined as:

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i \quad \text{where } \epsilon_{ij} \sim N(0, \sigma).$$

Consider the following hypotheses:

$$H_0 : \boldsymbol{\mu} \in C_0, \quad H_1 : \boldsymbol{\mu} \in C_1, \quad \text{and} \quad H_2 : \boldsymbol{\mu} \in C_2.$$

We want to test H_0 against $H_1 - H_0$ or H_1 against $H_2 - H_1$.

Notation: When we refer to test of H_0 against H_1 , it should be read as H_0 against $H_1 \setminus H_0$; in the literature, $H_1 \setminus H_0$ is also written as $H_1 - H_0$ [15].

Consider data in Table 5.1 below representing the size of pituitary fissure for a group of young children between ages 8-14 years.

Table 5.1: Size of Pituitary Fissure

Age	Size			n_i	$\bar{y}_i$
8	21	23.5	23	3	22.5
10	24	21	25	3	23.33
12	21.5	22	19	3	20.83
14	23.5	25		2	24.25

- i) If the size does not increase with age (H_1 is true), how can we estimate μ'_i 's?
- ii) Given MLEs under H_0 and H_1 , how can we find the null distribution of the likelihood

ratio test (LRT) statistic?

iii) How can we verify the presumption H_1 ?

The MLE under H_ℓ is the solution of $\min_{\boldsymbol{\mu} \in C_\ell} \sum_{i=1}^k w_i (\bar{y}_i - \mu_i)^2$ where $w_i = \frac{n_i}{\sigma^2}$ and $\ell = 1, 2, 3$.
Let $P_w(\bar{\mathbf{y}}|C_\ell)$ be the least square projection of $\bar{\mathbf{y}}$ onto C_ℓ . Therefore,

MLE of $\boldsymbol{\mu}$ under H_2 : $\hat{\boldsymbol{\mu}} = P_w(\bar{\mathbf{y}}|C_2) = (\bar{y}_1, \dots, \bar{y}_k)^T$

MLE of $\boldsymbol{\mu}$ under H_0 : $\bar{\boldsymbol{\mu}} = P_w(\bar{\mathbf{y}}|C_0) = (\bar{y}, \dots, \bar{y})^T$: vector of grand mean

MLE of $\boldsymbol{\mu}$ under H_1 : $\tilde{\boldsymbol{\mu}} = P_w(\bar{\mathbf{y}}|C_1)$: need some algorithms like GP

Figure 5.1: Geometry of constrained LRT

The LRT is given as

$$\text{LRT of } H_0 \text{ against } H_1 - H_0 : T_{01} = \sum_{i=1}^k w_i (\tilde{\mu}_i - \bar{\mu}_i)^2$$

$$\text{LRT of } H_1 \text{ against } H_2 - H_1 : T_{12} = \sum_{i=1}^k w_i (\hat{\mu}_i - \tilde{\mu}_i)^2$$

$$\text{LRT of } H_0 \text{ against } H_2 - H_0 : T_{02} = \sum_{i=1}^k w_i (\hat{\mu}_i - \bar{\mu}_i)^2$$

The null distribution of T_{02} under $H_0 : \mu_1 = \cdots = \mu_k$ is $\chi^2(k-1)$.

What is the distribution of T_{01} (or T_{12}) under H_0 ?

Note 5.1: Let $\|\mathbf{y} - \mathbf{x}\|_{\mathbf{w}}^2 = \sum_{i=1}^k w_i(y_i - x_i)^2$. We might be interested in $\min_{\mathbf{A}\boldsymbol{\mu} \leq \mathbf{0}} \|\bar{\mathbf{y}} - \boldsymbol{\mu}\|_{\mathbf{w}}^2$, where

- $\mathbf{A}$ is $m \times k$ constraints coefficient matrix and $\mathbf{A}\boldsymbol{\mu} \leq \mathbf{0}$ is the set of linear inequalities.

Example 5.2 (Multinomial): Let $\mathbf{X} = (X_1, \dots, X_k)^T \sim MN(n; \pi_1, \dots, \pi_k)$ follows a multinomial distribution.

Consider testing H_0 against $H_1 - H_0$ or H_1 against $H_2 - H_1$, where

$$H_0 : \pi_i = \frac{1}{k}, i = 1, \dots, k$$

$$H_1 : \pi_1 \leq \pi_2 \leq \cdots \leq \pi_k$$

$$H_2 : \text{no restriction.}$$

What is the MLE of $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)^T$ under H_1 ? Here the kernel likelihood is $\prod_{i=1}^k \pi_i^{n\hat{\pi}_i}$.

We have the maximization problem

$$\begin{aligned} & \underset{\boldsymbol{\pi}}{\text{maximize}} && \prod_{i=1}^k \pi_i^{n\hat{\pi}_i} \\ & \text{subject to} && \pi_1 \leq \pi_2 \leq \cdots \leq \pi_k \\ & && \sum_{i=1}^k \pi_i = 1 \end{aligned}$$

Here the restricted MLE is $\tilde{\boldsymbol{\pi}} = P_w(\hat{\boldsymbol{\pi}}|C_1)$, where $w_i = 1$, $i = 1, \dots, k$ and the LRT gives the test statistic

$$T_{01} = 2n \sum_{i=1}^k \hat{\pi}_i \left[\ln(\tilde{\pi}_i) - \ln\left(\frac{1}{k}\right) \right].$$

The asymptotic null distribution of T_{01} needs to be derived.

Example 5.3 (Multinomial): Let $\mathbf{X} = (X_1, \dots, X_k)^T \sim MN(n; \pi_1, \dots, \pi_k)$. Let $\mathbf{q} = (q_1, \dots, q_k)^T$ another probability vector. Consider testing H_0 against $H_1 - H_0$, where

$$H_0 : \boldsymbol{\pi} = \mathbf{q}$$

$$H_1 : \sum_{j=1}^i \pi_j \leq \sum_{j=1}^i q_j, \quad i = 1, 2, \dots, k-1$$

- (1) Assume $\mathbf{q}$ is known (one sample case). What is the MLE of $\boldsymbol{\pi}$ under H_1 ?

We have the maximization problem

$$\begin{aligned} & \underset{\boldsymbol{\pi}}{\text{maximize}} && \prod_{i=1}^k \pi_i^{n\hat{\pi}_i} \\ & \text{subject to} && \sum_{j=1}^i \pi_j \leq \sum_{j=1}^i q_j, \quad i = 1, 2, \dots, k-1 \\ & && \sum_{i=1}^k \pi_i = 1. \end{aligned}$$

Here the restricted MLE is $\tilde{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}} P_{\hat{\boldsymbol{\pi}}} \left(\frac{\mathbf{q}}{\hat{\boldsymbol{\pi}}} | C_1 \right)$ where $\frac{\mathbf{q}}{\hat{\boldsymbol{\pi}}} = \left(\frac{q_1}{\hat{\pi}_1}, \dots, \frac{q_k}{\hat{\pi}_k} \right)^T$ and the LRT

$$T_{01} = 2n \sum_{i=1}^k \hat{\pi}_i [\ln(\tilde{\pi}_i) - \ln(q_i)].$$

The asymptotic null distribution of T_{01} needs to be derived.

- (2) Given an additional sample $\mathbf{Y} = (Y_1, \dots, Y_k)^T \sim MN(m; q_1, \dots, q_k)$,

- the unrestricted MLE for $\boldsymbol{\pi}$ and $\mathbf{q}$ are $\hat{\boldsymbol{\pi}} = \frac{\mathbf{X}}{n}$ and $\hat{\mathbf{q}} = \frac{\mathbf{Y}}{m}$, respectively,
- the restricted MLE for $\boldsymbol{\pi}$ and $\mathbf{q}$ are $\tilde{\boldsymbol{\pi}}$ and $\tilde{\mathbf{q}}$, which need to be obtained, and
- the null distribution of the LRT T_{01} needs to be derived.

5.2 Comparison of Population Order Means

To illustrate the one-way ANOVA test with constraints, this section leverages ideas and methodology from [15].

Using the same concept as the one-way ANOVA model, we will consider order restrictions on this model to demonstrate constrained inference. Suppose that there are k treatments to be compared. Let y_{ij} denote the j^{th} observation for Treatment i , (see Table 5.2). Let

$$y_{ij} = \mu_i + \epsilon_{ij} \quad \text{with } i = 1, \dots, k \text{ and } j = 1, \dots, n_i, \quad (5.1)$$

where

- * y_{ij} are mutually independent,
- * μ_i is the location parameter for treatment i , and
- * σ^2 is the common variance.

Table 5.2: Comparison of k means

Treatment	Independent Observation	Sample Mean	Population Distribution (cdf)
1	$y_{11}, \dots, y_{1n_1}$	$\bar{y}_1$	$F\left(\frac{t-\mu_1}{\sigma^2}\right)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$y_{1k}, \dots, y_{kn_k}$	$\bar{y}_y$	$F\left(\frac{t-\mu_k}{\sigma^2}\right)$

Let the null and alternative hypotheses be:

$$H_0 : \mu_1 = \dots = \mu_k \quad \text{vs} \quad H_1 : \mu_i - \mu_\ell \geq 0 \text{ for } i, \ell = 1, \dots, k. \quad (5.2)$$

Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^T$ and H can be H_0, H_1 or H_2 , where

$$H_2 : \mu_1, \dots, \mu_k \text{ are not restricted.} \quad (5.3)$$

Define the *Residual Sum Square (RSS)* under H as

$$RSS(H) = \inf_{\boldsymbol{\mu} \in H} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2. \quad (5.4)$$

With this definition of RSS, we can clearly see that it leads to

$$RSS(H_0) = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2, \quad \text{and} \quad RSS(H_2) = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \quad (5.5)$$

When testing H_0 against the restricted hypothesis H_1 we can use the F -test in (5.6) for H_0 against H_2 on $(2, \nu)$ degrees of freedom, where ν is the error degrees of freedom. This is possible because we start off with the same null hypothesis in both instances. We note, however, that the standard F -test for H_0 against H_2 is **not expected** to have good power properties for testing H_0 against H_1 ; this is a result of not using the added restriction of $\mu_i \geq \mu_\ell$, which means that the test is not set up to detect departures in the direction of H_1 . Recall the definition of the standard F -statistic:

$$F = \frac{\{RSS(H_0) - RSS(H_2)\}(k-1)^{-1}}{S^2}, \quad (5.6)$$

where

- $S^2 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n-k}$ is the error mean square, which is the unbiased estimate of the common variance σ^2 , with the degree of freedom $\nu = n - k$ and $n = \sum_{i=1}^k n_i$,
- the distribution of S^2 is $\frac{\sigma^2 \chi_\nu^2}{\nu}$ which is independent from $\bar{y}_1, \dots, \bar{y}_k$,
- $\bar{y}$ is the grand mean,
- the numerator of F is a measure of the discrepancy between H_0 and H_2 , and
- the denominator, S^2 , acts as a scaling factor so that the null distribution of the test statistic does not depend on the unknown scale parameter σ for the error.

Remark 5.1 (RSS for Alternative Hypothesis): *Technically, $RSS(H_2)$ should be written as $RSS(H_0 \cup H_2)$, but since H_0 is on the boundary of H_2 , the value of $RSS(H_2)$ is not affected.*

Given the above information, we can test H_0 against H_1 by obtaining a reasonable test statistic

through the modification of the F -statistic in the following manner:

$$\bar{F} = \frac{RSS(H_0) - RSS(H_1)}{S^2}, \quad (5.7)$$

where

* $RSS(H_1) = \min_{\boldsymbol{\mu} \in H_1} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \tilde{\mu}_i)^2$ is sum of squares of the residuals under H_1 ,

* $\tilde{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\mu} \in H_1} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$ is the point at which the sum of squares $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$ is minimized subject to the constraint in H_1 ,

* $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_k)^T$ is a restricted estimate of $(\mu_1, \dots, \mu_k)^T$ under H_1 .

* To compute the restricted estimator $\tilde{\boldsymbol{\mu}}$ under H_1 , it is sufficient to minimize $\sum_{i=1}^k n_i (\bar{y}_i - \mu_i)^2$ since

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k n_i (\bar{y}_i - \mu_i)^2.$$

* If the errors are identically independently distributed (iid) as $N(0, \sigma^2)$, it can be shown that $\tilde{\boldsymbol{\mu}}$ is the MLE of $\boldsymbol{\mu}$ under H_1 .

* The numerator of $\bar{F}$ is a measure of the discrepancy between H_0 and the restricted alternative H_1 .

$\bar{F}$ -test is simple to use/understand since it relies on the same principle as the standard F -test while also including additional information as in H_1 .

5.2.1 Computing Restricted F and E Test

The implementation of the restricted $\bar{F}$ and $\bar{E}^2$ -tests requires the computation of $RSS(H_1)$ that is equal to

$$RSS(H_1) = \min_{\boldsymbol{\mu} \in H_1} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2, \text{ where } H_1 : \mu_i - \mu_\ell \geq 0.$$

Let

$$q(\boldsymbol{\mu}) = (\bar{\mathbf{y}} - \boldsymbol{\mu})^T \mathbf{D}(\bar{\mathbf{y}} - \boldsymbol{\mu}), \quad (5.8)$$

where

$$\bullet \boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^T, \bar{\boldsymbol{\mu}} = (\bar{y}_1, \dots, \bar{y}_k)^T \text{ and } \mathbf{D} = \text{diag}\{n_1, \dots, n_k\}.$$

Let $\mathbf{A}$ be a matrix, where each row of $\mathbf{A}$ is a permutation of the k -vector $(1, -1, 0, \dots, 0)$, such that

$$\{\boldsymbol{\mu} : \mu_i - \mu_\ell \geq 0 \quad \forall i, \ell = 1, \dots, k\} = \{\boldsymbol{\mu} : \mathbf{A}\boldsymbol{\mu} \geq 0\}.$$

Since

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = q(\boldsymbol{\mu}) + C(\bar{\mathbf{y}}),$$

where $C(\bar{\mathbf{y}})$ does not depend on $\boldsymbol{\mu}$, we have

$$\begin{aligned} \bar{F} &= \frac{\min_{H_0} q(\boldsymbol{\mu}) - \min_{H_1} q(\boldsymbol{\mu})}{S^2}, \\ \bar{E}^2 &= \frac{\min_{H_0} q(\boldsymbol{\mu}) - \min_{H_1} q(\boldsymbol{\mu})}{\min_{H_0} q(\boldsymbol{\mu}) + C(\bar{\mathbf{y}})}. \end{aligned}$$

This constrained minimization problem in which the objective function $q(\boldsymbol{\mu})$ is a quadratic in $\boldsymbol{\mu}$ and the constraints are linear equality and inequality constraints in $\boldsymbol{\mu}$ is called a **quadratic program**. The $\bar{E}^2$ -test rejects H_0 for large values of $\bar{E}^2$. If the error distribution is **normal**, then it may be verified that

$$\bar{E}^2 = \frac{RSS(H_0) - RSS(H_1)}{RSS(H_0)} = \{1 - \exp(-LRT/n)\},$$

where LRT denotes the likelihood ratio statistic ($= -2 \log \Lambda$) for testing H_0 against H_1 .

5.2.2 The Null Distribution of Restricted F-Test when $k=3$

Given only three treatments $k = 3$ and the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$, we consider this a *special case* where we have three possible ordered alternatives, which are relevant to one-way ANOVA:

Table 5.3: Ordered Alternatives and ρ

H_1	ρ
$\mu_1 \leq \mu_2 \leq \mu_3$	$-\sqrt{\frac{n_1 n_3}{(n_1 + n_3)(n_2 + n_3)}}$
$\mu_1 \leq \mu_2$ and $\mu_1 \leq \mu_3$	$\sqrt{\frac{n_2 n_3}{(n_1 + n_2)(n_1 + n_3)}}$
$\mu_1 \leq \mu_2$	1 [i.e. $(w_0, w_1, w_2) = (0, 0.5, 0.5)$]

When testing H_0 against any of the three order restrictions, H_1 , shown in Table 5.3, the $\bar{F}$ null distribution is given by:

$$P(\bar{F} \leq c | H_0) = w_0 + w_1 P(F_{1,\nu} \leq c) + w_2 P(2F_{2,\nu} \leq c) \quad \text{with } (c > 0), \quad (5.9)$$

where the weights will be computed as:

$$w_1 = 0.5, \quad w_2 = (0.5 - \kappa), \quad \text{with } \kappa = (2\pi)^{-1} \cos^{-1}(\rho), \quad \text{and } w_0 + w_1 + w_2 = 1. \quad (5.10)$$

We choose notation $\bar{F}$ because of its relation to the unrestricted F -ratio, and also because its null distribution is a weighted average of probabilities associated with F -distributions. The p -value for the $\bar{F}$ -test is given by:

$$p\text{-value} = w_1 P(F_{1,\nu} \geq \bar{f}_{obs}) + w_2 P(2F_{2,\nu} \geq \bar{f}_{obs}), \quad (5.11)$$

where $\bar{f}_{obs}$ is the sample value of $\bar{F}$.

5.2.3 The Null Distribution of Restricted F when k is more than 3

Let's consider another case when we have more than three treatments. The theorem below states that the null distribution does not depend on (cdf F, μ, σ) of the error distribution.

Theorem 5.1: When finding the sampling distribution of the restricted F -test, $\bar{F}$ does not depend on the common value μ of $\mu_1, \dots, \mu_k$ under H_0 and the common variance σ^2 , but it does depend on the functional form of cdf F in Table (5.2). Also, when the limit is taken as $n = \sum n_i \rightarrow \infty$, the asymptotic null distribution of $\bar{F}$ does not depend on (cdf F, μ, σ).

Proof. Let $y_{ij}^* = \frac{y_{ij} - \mu}{\sigma} \quad \forall (i, j)$, then

$$RSS^*(H_0) = \sigma^{-2}RSS(H_0), \quad RSS^*(H_1) = \sigma^{-2}RSS(H_1), \quad \text{and } (S^*)^2 = \sigma^{-2}S^2,$$

and hence

$$\bar{F} = \frac{RSS(H_0) - RSS(H_1)}{S^2} = \frac{RSS^*(H_0) - RSS^*(H_1)}{(S^*)^2} = \bar{F}^*.$$

Under H_0 , the distribution of y_{ij}^* follows the cdf $F(t)$, which does not depend on (μ, σ) . Now, since $\bar{F}^*$ is a function of y_{ij}^* only and $\bar{F} = \bar{F}^*$, it follows that the distribution of $\bar{F}$ does not depend on (μ, σ) . $\square$

5.2.3.1 Computation of the exact p-value for the restricted F test

Table 5.2 provides the functional form F , (cdf), of the error distribution. Eq. (5.2) defines the testing problem H_0 against H_1 and Eq. (5.7) defines the test statistic $\bar{F}$. Given this information, we can use simulation by following the next steps to find the p -value for the restricted F -test, $\bar{F}$:

- (1) Generate independent observations $\{y_{ij} : i = 1, \dots, k, j = 1, \dots, n_i\}$ from the cdf $F\left(\frac{t - \mu_0}{\sigma_0}\right)$ where (μ_0, σ_0) can have any values, but must be held fixed for different (i, j) .

Note: since Theorem 5.1 states that the null distribution of $\bar{F}$ does not depend on the common μ or σ , we can generate the observations from a distribution with any value for the common location and scale parameters.

- (2) Compute the test statistic $\bar{F}$ in Eq. (5.7) with $RSS(H_0)$ and $RSS(H_1)$ as in Eq.(5.5) and Eq.(5.7), respectively.
- (3) Repeat steps (1) and (2), above, N times (here, we will use $N = 10000$). Then estimate the p -value using $\frac{M}{N}$, where M is the number of times the $\bar{F}$ statistic in step (2) was greater than its sample value $\bar{f}_{obs}$.

Remark 5.2: The second part of Theorem 5.1 says the asymptotic distribution of restricted F -test, $\bar{F}$, does not depend on the error distribution (F, μ, σ) , which may be unknown. The simulation method used to estimate the asymptotic p -value when the error distribution is normal can also be used for any error distribution when n_i is large for $i = l, \dots, k$.

If the precise form of cdf F is unknown, but we know that cdf F is a member of the class of $\mathcal{F}$ distributions for some $\mathcal{F}$, then let p_F represent the p -value corresponding to cdf F . We observe:

$$p\text{-value} = \sup_{F \in \mathcal{F}} p_F. \quad (5.12)$$

For example, suppose we know that the error distribution is normal, logistic, or a t-distribution with four degrees of freedom (T_4), then we can obtain the p -value in Eq.(5.12) for $\bar{F}$ by using the previous simulation method to calculate the p -values corresponding to normal, logistic, and T_4 and then using their maximum as the p -value in Eq.(5.12).

Remark 5.3 (General Remarks): If the errors ϵ_{ij} are independent and distributed as $N(0, \sigma^2)$, then the p -value for $\bar{F}$ is:

$$\sum_{i=1}^k w_i(H_0, H_1) P(iF_{i,\nu} \geq \bar{f}_{obs}),$$

where

- $\{w_i(H_0, H_1)\}$ are quantities known as **chi-bar-square** weights and also as level probabilities. They are non-negative weights, which depend on the null hypothesis H_0 and the alternative hypothesis H_1 , and
- $\bar{f}_{obs}$ is the sample value of $\bar{F}$.

Example 5.4 (Ordered Treatment Means in One-Way Layout): The experiment described in [15] evaluates the impact of certain exercises on the age at which a child starts to walk. The data in Table 5.4 provides information on Y , which represents the age (in months) at which a child starts to walk.

Table 5.4: The age at which a child first walks

Treatment (i)		Age (in months)					n_i	$\bar{y}_i$	μ_i
1	9.00	9.50	9.75	10.00	13.00	9.50	6	10.125	μ_1
2	11.00	10.00	10.00	11.75	10.50	15.00	6	11.375	μ_2
3	13.25	11.50	12.00	13.50	11.50		5	12.35	μ_3
4	11.50	12.00	9.00	11.50	13.25	13.00	6	11.7	μ_4

- Treatment group 1 completed a special 12 minutes per day walking exercise, beginning at age 1 week and lasting 7 weeks.
- Treatment group 2 completed daily exercises, but not the special walking exercises.
- Treatment group 3 is the control; they did not receive any exercises or other treatments.
- Treatment group 4 did not receive any special exercises, but were monitored weekly for progress.

For Treatment i ($i = 1, 2, 3, 4$), let

μ_i = Mean age (in months) at which a child starts to walk.

The traditional ANOVA test is:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{versus} \quad H_2 : \mu_1, \mu_2, \mu_3, \text{ and } \mu_4 \text{ are not all equal.}$$

In our example, we want to incorporate additional information. Suppose the researcher assumed that the walking exercises had no negative impact on the mean age at which a child starts to walk. We would like to include this information to improve our statistical analysis. To illustrate this, we assume the researcher wants to incorporate the following information (mean order restriction): $\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$. In this case, the testing problem is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ versus } H_1 : \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \text{ and } \mu_1, \mu_2, \mu_3, \text{ and } \mu_4 \text{ are not all equal.}$$

This is equivalent to

$$H_0 : \boldsymbol{\mu} \in C_0 \text{ versus } H_1 : \boldsymbol{\mu} \in C_1.$$

The traditional ANOVA, where we test H_0 against H_2 , fails to include the additional information we need. So we can do better than the traditional F -test by using the restricted F -test, $\bar{F}$

For simplicity, consider only 3 treatments. In this case, the testing problem is

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ versus } H_1 : \mu_1 \leq \mu_2 \leq \mu_3.$$

If we minimize $\sum_{i=1}^3 n_i(\bar{y}_i - \mu_i)$, the unrestricted estimate of $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ is $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, \bar{y}_3) = (10.125, 11.375, 12.35)$, which is also the restricted estimate $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3) = \bar{\mathbf{y}}$, since $\bar{\mathbf{y}}$ satisfies the constraints in H_1 , it follows that the estimate of $\boldsymbol{\mu}$ subject to the constraint in H_1 is also equal to the unrestricted estimate $\bar{\mathbf{y}}$. The restricted $\bar{F}$ -test sample value is

$$\bar{F} = \frac{RSS(H_0) - RSS(H_1)}{S^2} = \frac{45.927 - 32.137}{2.296} = 5.978.$$

Since the cdf F follows a normal distribution with $\mu = 0$ and σ , which is unknown, then for testing H_0 against $H_1 : \mu_1 \leq \mu_2 \leq \mu_3$, the p -value can be computed using the equation (5.11) and $\rho = -0.5$, that is,

$$p\text{-value} = 0.5P(F_{1,14} \geq 5.978) + 0.17P(F_{2,14} \geq 2.989) = 0.028.$$

For testing

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{against} \quad H_2 : \mu_1, \mu_2, \mu_3 \text{ are not all equal,}$$

the p -value for the unrestricted F -statistic $= \frac{(k-1)^{-1}(RSS(H_0) - RSS(H_1))}{S^2} = 2.989$ is

$$p\text{-value} = P(F_{2,14} \geq 2.989) = 0.083.$$

We can see that the restricted p -value is **smaller** than the unrestricted p -value. We can comfortably expect the $\bar{F}$ -test (testing H_0 against H_1) to provide stronger evidence to reject H_0 than the unrestricted F -test when the sample means satisfy the order $\bar{y}_1 \leq \bar{y}_2 \leq \bar{y}_3$. When the number of order restrictions is four or more, then the null distribution of $\bar{F}$ is a weighted sum similar to (5.9); however, it is usually rather inconvenient to use it for computing the exact p -value. This is why a **simulation** approach offers a simple and practical method of computing a sufficiently precise p -value no matter the number of order restrictions, for any error distribution.

Now, consider 4 or more treatments.

Suppose we want to test H_0 against an alternative H_1 that incorporates our previous information. There is no single way to formulate the alternative hypothesis in this situation, where:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{against} \quad H_1 : \mu_1 \leq \mu_3, \mu_2 \leq \mu_3, \mu_1 \leq \mu_4, \mu_2 \leq \mu_4.$$

Here, we see a common characteristic of the majority of the problems related to tests against inequality constraints: there is no convenient formula to obtain the p -value for $\bar{F}$. Since $\bar{y}$, the vector of sample means, satisfies the restriction in H_1 , the constrained estimator $\tilde{\boldsymbol{\mu}}$ and unconstrained estimator are the same, where $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ is $\bar{\mathbf{y}} = (10.1, 11.4, 12.4, 11.7)$. The restricted $\bar{F}$ -test sample value is

$$\bar{F} = \frac{RSS(H_0) - RSS(H_1)}{S^2} = \frac{58.46739 - 43.68958}{2.299452} = 6.43.$$

The simulation approach was used to obtain the p -values corresponding to a range of error distributions. The p -value in the last column was computed by re-sampling with replacement

Table 5.5: The p -values for the $\bar{F}$ -test for different error distributions

Test	$N(0, \sigma)$	T_4	T_{10}	χ_1^2	χ_2^2	χ_4^2	χ_7^2	RS
$\bar{F}$ - [15]	0.052	0.051	0.058	0.050	0.048	0.049	0.051	0.048
$\bar{F}$ -Thesis	0.049	0.044	0.051	0.037	0.044	0.045	0.047	0.044

from the error distribution, where the error distribution is the empirical distribution of the residuals about the treatment means. The p -values in Table (5.5) are close for different error distributions; the p -values in the first row are given by [15], and the p -values in the second row are replicated values. The simulation method illustrates a convenient way of implementing tests against any order restriction even when the errors are iid and the common distribution is not normal. The unconstrained F -statistic for testing H_0 against H_2 is 2.14 and its p -value is 0.129 based on the F -statistic that follows the F -distribution with degree of freedoms 3 and 19. If the sample means satisfy the constraints that correspond to those in the alternative hypothesis, then the estimate of $\boldsymbol{\mu}$ under H_0 and H_1 are the same and the p -value for the constrained test would be smaller than that for the unconstrained F -test.

5.3 Constrained Tests on Multivariate Normal Mean

In multivariate analysis, for a given $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$, $\boldsymbol{\theta}$ and $\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}$ we normally define constraints imposed on model parameters in terms of linear equality constraints (i.e. $\mathbf{A}_1\boldsymbol{\theta} = \mathbf{b}_1$) and inequality constraints (i.e. $\mathbf{A}_2\boldsymbol{\theta} \leq \mathbf{b}_2$). This section covers estimation and testing procedures for these equality and inequality constraints. When conducting standard hypothesis testing, where we test $H_0 : \mathbf{A}\boldsymbol{\theta} = \mathbf{0}$ against $H_a : \mathbf{A}\boldsymbol{\theta} \neq \mathbf{0}$, and where $\mathbf{A}$ is a given fixed matrix when observations are iid from the multivariate normal distribution, $N_p(\boldsymbol{\theta}, \mathbf{V})$, we can easily apply the LRT. This is possible since we can compute the LRT statistic without much

difficulty and the statistical tables for its null distribution are easily available, where the null distribution is χ_q^2 with $q = \text{rank}(\mathbf{A})$. This theory becomes more complicated if the hypotheses contain inequalities in $\boldsymbol{\theta}$. Furthermore, we cannot apply the results without handling several difficulties, which include the fact that the null distribution of LRT depends on the matrix $\mathbf{A}$ through $\mathbf{A}\mathbf{V}\mathbf{A}^T$, not just on $\text{rank}(\mathbf{A})$. Another issue is that it is extremely difficult to exactly calculate the critical values. Simulation helps resolve this issue because we can use simulation to compute the p -values and critical values of the tests mentioned earlier in this section.

5.3.1 Likelihood Function

When the population distribution is $N_p(\boldsymbol{\theta}, \mathbf{V})$, the nature of the solutions to inequality constrained testing problems also depend on what is known about the variance-covariance matrix $\mathbf{V}$ of order p in addition to the structure of the null and alternative parameter spaces. Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ be n iid observations from $N_p(\boldsymbol{\theta}, \mathbf{V})$, where $\mathbf{V}$ is a positive definite matrix of order p and $\boldsymbol{\theta} \in \mathbb{R}^p$. The log-likelihood for the n observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ is

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\frac{2}{n} \log |\mathbf{V}| - (np/2) \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Y}_i - \boldsymbol{\theta}) \\ &= \ell(\boldsymbol{\theta}) + g(\mathbf{Y}_1, \dots, \mathbf{Y}_n, \mathbf{V}), \end{aligned}$$

where $g(\mathbf{Y}_1, \dots, \mathbf{Y}_n, \mathbf{V})$ does not depend on $\boldsymbol{\theta}$ and $\ell(\boldsymbol{\theta})$ is the kernel of the log-likelihood for $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, given by

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2} (\bar{\mathbf{Y}} - \boldsymbol{\theta})^T (n^{-1} \mathbf{V})^{-1} (\bar{\mathbf{Y}} - \boldsymbol{\theta}) = -\frac{1}{2} \|\bar{\mathbf{Y}} - \boldsymbol{\theta}\|_{\mathbf{V}}^2.$$

Since the distribution for the mean vector $\bar{\mathbf{Y}}$ is normally distributed with mean $\boldsymbol{\theta}$ and variance-covariance $n^{-1} \mathbf{V}$ i.e. $\bar{\mathbf{Y}} \sim N_p(\boldsymbol{\theta}, n^{-1} \mathbf{V})$, the kernel of the log-likelihood for a single observation $\bar{\mathbf{Y}}$ is $\ell(\boldsymbol{\theta})$. Therefore, the MLE and the LRT based on $\bar{\mathbf{Y}}$ and those based on $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are the same [15].

5.3.2 Constrained MLE and LRT

To find the constrained MLE and to derive the null distribution of the constrained LRT when the alternative hypothesis involves inequality constraints, we start with the simple special cases of standard basis (i.e. the two orthogonal unit vectors pointing in the direction of the axes of a Cartesian coordinate system) and non-standard basis (i.e. the two linear independent unit vectors representing an angle ω rotation of the 2-D standard basis). We then use these ideas to introduce the general estimation and testing results.

Example 5.5 (MLE in Two Dimensions): Consider a simple bivariate normal $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2(\boldsymbol{\theta}, \mathbf{I})$, where $\mathbf{I}$ is the 2×2 identity matrix and $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$. Consider the maximum likelihood estimation of $\boldsymbol{\theta}$ based on a single observation of $\mathbf{Y}$, and subject to the constraint $\boldsymbol{\theta} \in \mathcal{C}$, where $\mathcal{C} = \{\boldsymbol{\theta} : \mathbf{A}\boldsymbol{\theta} \geq \mathbf{0}\}$ is a closed convex cone formed by the two rows of $\mathbf{a}_1^T$ and $\mathbf{a}_2^T$ of the 2×2 nonsingular matrix $\mathbf{A}$, and $\mathcal{C}^0 = \{\boldsymbol{\alpha} : \boldsymbol{\alpha}^T \boldsymbol{\theta} \leq 0 \quad \forall \boldsymbol{\theta} \in \mathcal{C}\}$ is the negative dual or polar cone of $\mathcal{C}$ w.r.t. the inner product $\boldsymbol{\alpha}^T \boldsymbol{\theta} = \alpha_1 \theta_1 + \alpha_2 \theta_2$. The boundaries of $\mathcal{C}^0$ are the orthogonals to the boundaries of $\mathcal{C}$, and that $\mathcal{C}^0$ is the closed convex cone formed by these orthogonals (see Figure 5.2a). It is clear that $\mathcal{C}$ and $\mathcal{C}^0$ partition the plane into 4 cones denoted as $\{\mathcal{S}_1 = \mathcal{C}, \mathcal{S}_2, \mathcal{S}_3 = \mathcal{C}^0, \mathcal{S}_4\}$. Let $\mathbf{u}$ and $\mathbf{v}$ be unit vectors parallel to the upper and lower boundaries of $\mathcal{C}$.

For the single observation $\mathbf{Y}$, the kernel $\ell(\boldsymbol{\theta})$ of the log-likelihood is given by

$$-2\ell(\boldsymbol{\theta}) = \|\mathbf{Y} - \boldsymbol{\theta}\|^2 = \{(Y_1 - \theta_1)^2 + (Y_2 - \theta_2)^2\}.$$

Let $\tilde{\boldsymbol{\theta}}$ be the constrained MLE of $\boldsymbol{\theta}$ subject to $\mathbf{A}\boldsymbol{\theta} \geq \mathbf{0}$. Since $-2\ell(\boldsymbol{\theta})$ is equal to the squared distance between $\mathbf{Y}$ and $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$ is the point in $\mathcal{C}$ that is closest to $\mathbf{Y}$. In other words, $\tilde{\boldsymbol{\theta}}$ is the projection of $\mathbf{Y}$ onto $\mathcal{C}$ (see Figure 5.2a). Let P denote the projection function, we can write

$$\tilde{\boldsymbol{\theta}} = P(\mathbf{Y}|\mathcal{C}).$$

Figure 5.2: Two dimensions constrained MLE of θ subject to $\mathbf{A}\theta \geq \mathbf{0}$, and the LRT of H_0 vs H_1 and a typical boundary of the critical region is ABCD, based on a single observation of $\mathbf{Y}$, where $\mathbf{Y} \sim N(\theta, \mathbf{I})$ [15].

Then the constrained MLE $\tilde{\theta}$ is given by

$$\tilde{\theta} = P(\theta|\mathcal{C}) = \begin{cases} \mathbf{Y} & \text{if } \mathbf{Y} \in \mathcal{S}_1 \\ (\mathbf{u}^T \mathbf{Y}) \mathbf{u} & \text{if } \mathbf{Y} \in \mathcal{S}_2 \\ \mathbf{0} & \text{if } \mathbf{Y} \in \mathcal{S}_3 \\ (\mathbf{v}^T \mathbf{Y}) \mathbf{v} & \text{if } \mathbf{Y} \in \mathcal{S}_4 \end{cases}$$

Since $\tilde{\theta}$ is a function of $\mathbf{Y}$ only, and we know the distribution function of $\mathbf{Y}$, we can extrapolate explicit expressions for the distribution of $\tilde{\theta}$. In comparison to the parameter space for θ , which is $\mathbb{R}^2$, we note that the distribution of the constrained MLE, $\tilde{\theta}$, is not normal. Therefore, we cannot use conventional methods to find the confidence region for θ based on the distribution of $(\tilde{\theta} - \theta)$. As a result, $(\tilde{\theta} - \theta)$ does not have as much success in statistical inference. Consider the LRT of:

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ vs } H_1 : \mathbf{A}\boldsymbol{\theta} \geq \mathbf{0}$$

based on a single observation of $\mathbf{Y}$, since $-2\ell(\boldsymbol{\theta}) = \|\mathbf{Y} - \boldsymbol{\theta}\|^2$ and

$$LRT = 2[\max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in H_1\} - \max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in H_0\}],$$

the LRT statistic is

$$LRT = \|\mathbf{Y}\|^2 - \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|^2 = \mathbf{Y}^T \mathbf{Y} - (\mathbf{Y} - \tilde{\boldsymbol{\theta}})^T \mathbf{Y} + (\mathbf{Y} - \tilde{\boldsymbol{\theta}})^T \tilde{\boldsymbol{\theta}} = \|\tilde{\boldsymbol{\theta}}\|^2.$$

We can verify that $(\mathbf{Y} - \tilde{\boldsymbol{\theta}})^T \tilde{\boldsymbol{\theta}} = 0$ using Figure (5.2a) and considering the value of $\tilde{\boldsymbol{\theta}}$ in each of the four cases $\mathcal{S}_i$, $i = 1, 2, 3, 4$ when $\tilde{\boldsymbol{\theta}}$ is not zero (i.e. $\mathbf{Y} \notin \mathcal{S}_3$). Our focus is on the distribution of the LRT under null hypothesis, so we suppose the null hypothesis to be true for the remainder of these derivations. We obtain the expression for $pr(LRT \leq c)$ by:

$$pr(LRT \leq c) = \sum_{i=1}^4 pr(\{LRT \leq c\} \cap \{\mathbf{Y} \in \mathcal{S}_i\}) = \sum_{i=1}^4 pr(LRT \leq c | \mathbf{Y} \in \mathcal{S}_i) pr(\mathbf{Y} \in \mathcal{S}_i)$$

Let's evaluate each of the conditional probabilities in the last expression. The conditional distribution of $Y_1^2 + Y_2^2$, given that the direction of $\mathbf{Y}$ is in $\mathcal{S}_1$, is the same as that of its unconditional distribution, i.e. the length ($\|\mathbf{Y}\|^2 \leq c$) and direction of $\mathbf{Y}$ ($\mathbf{Y} \in \mathcal{S}_1$) are independent; refer to [34] page 279. Therefore,

$$\begin{aligned} pr(LRT \leq c | \mathbf{Y} \in \mathcal{S}_1) &= pr(Y_1^2 + Y_2^2 \leq c | \mathbf{Y} \in \mathcal{S}_1) \\ &= pr(Y_1^2 + Y_2^2 \leq c) = pr(\chi_2^2 \leq c). \end{aligned}$$

The conditional distribution of Y_2^2 , given that the direction of $\mathbf{Y} \in \mathcal{S}_2$, and using the new orthogonal coordinate system with OA and OD as the first and second axes, respectively, is

obtained as

$$\begin{aligned}
pr(LRT \leq c \mid \mathbf{Y} \in \mathcal{S}_2) &= pr(Y_2^2 \leq c \mid Y_2 \geq 0, Y_1 \leq 0), \\
&= pr(Y_2^2 \leq c \mid Y_2 \geq 0), \text{ since } Y_1 \text{ and } Y_2^2 \text{ are independent,} \\
&= pr(Y_2^2 \leq c), \text{ since } Y_2 \text{ is symmetric,} \\
&= pr(\chi_1^2 \leq c), \text{ since } Y_2 \sim N(0, 1).
\end{aligned}$$

Similarly, $pr(LRT \leq c \mid \mathbf{Y} \in \mathcal{S}_4) = pr(\chi_1^2 \leq c)$. Therefore, we have that

$$LRT = \begin{cases} Y_1^2 + Y_2^2 & \text{given } \mathbf{Y} \in \mathcal{S}_1 \\ (\mathbf{u}^T \mathbf{Y})^2 & \text{given } \mathbf{Y} \in \mathcal{S}_2 \\ 0 & \text{given } \mathbf{Y} \in \mathcal{S}_3 \\ (\mathbf{v}^T \mathbf{Y})^2 & \text{given } \mathbf{Y} \in \mathcal{S}_4 \end{cases} \sim \begin{cases} \chi_2^2 \\ \chi_1^2 \\ \chi_0^2 = 0 \\ \chi_1^2 \end{cases}$$

To maintain notational consistency, chi-square with zero degree of freedom takes a value of zero with the probability of one, i.e. $pr(\chi_0^2 \leq c) = 1$. From this, we see that the null distribution of the LRT is the weighted sum of chi-square distributions for $c > 0$:

$$\begin{aligned}
pr(LRT \leq c \mid H_0) &= w_0 pr(\chi_0^2 \leq c) + 0.5 pr(\chi_1^2 \leq c) + (0.5 - w_0) pr(\chi_2^2 \leq c) \\
&= \sum_{i=0}^2 w_i pr(\chi_i^2 \leq c),
\end{aligned}$$

where (w_0, w_1, w_2) are the probabilities that $\mathbf{Y}$ falls in the cones $\mathcal{S}_3$, $\mathcal{S}_2 \cup \mathcal{S}_4$, and $\mathcal{S}_1$, respectively, with $w_0 = pr(\mathbf{Y} \in \mathcal{S}_3 \mid H_0) = pr(\mathbf{Y} \in \mathcal{C}^0 \mid H_0) = (2\pi)^{-1} \gamma$ and γ is the angle (in radians) of the cone $\mathcal{C}^0$ at its vertex. Therefore,

$$w_0 = (2\pi)^{-1} \arccos \left[\mathbf{a}_1^T \mathbf{a}_2 / \sqrt{(\mathbf{a}_1^T \mathbf{a}_1)(\mathbf{a}_2^T \mathbf{a}_2)} \right].$$

The critical region, $\{LRT \geq c\}$, is the region to the upper right of the curve ABCD in Figure 5.2b; AB is orthogonal to the upper boundary of $\mathcal{C}$, CD is orthogonal to the lower boundary of $\mathcal{C}$ and BC is a circular arc of radius $\sqrt{c}$.

Remark 5.4: In the case of standard basis and the estimation subject to the constraint $\boldsymbol{\theta} \in \mathcal{C}$ where $\mathcal{C}$ is the nonnegative orthant $\{\boldsymbol{\theta} : \theta_1 \geq 0, \theta_2 \geq 0\}$, the four cones from above would be the four quadrants $\mathcal{Q}_i$, $i = 1, 2, 3, 4$ in the 2-D plane as seen in Figure 5.2a.

Figure 5.3: Two dimensions constrained MLE of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \geq \mathbf{0}$, and the LRT of H_0 vs H_1 and a typical boundary of the critical region is ABCD, based on a single observation $\mathbf{Y}$, where $\mathbf{Y} \sim N(\boldsymbol{\theta}, \mathbf{I})$.

Then the constrained MLE $\tilde{\boldsymbol{\theta}}$ is given by

$$\tilde{\boldsymbol{\theta}} = P(\boldsymbol{\theta} | \mathbb{R}^{+2}) = (\tilde{\theta}_1, \tilde{\theta}_2) = \begin{cases} (Y_1, Y_2) & \text{if } \mathbf{Y} \in \mathcal{Q}_1 \\ (0, Y_2) & \text{if } \mathbf{Y} \in \mathcal{Q}_2 \\ (0, 0) & \text{if } \mathbf{Y} \in \mathcal{Q}_3 \\ (Y_1, 0) & \text{if } \mathbf{Y} \in \mathcal{Q}_4 \end{cases}$$

and the likelihood ratio test (LRT) of $H_0 : \theta_1 = \theta_2 = 0$ vs $H_1 : \theta_1 \geq 0, \theta_2 \geq 0$ is

$$LRT = \|\mathbf{Y}\|^2 - \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|^2 = \|\tilde{\boldsymbol{\theta}}\|^2,$$

where

$$LRT = \begin{cases} Y_1^2 + Y_2^2 & \text{given } \mathbf{Y} \in \mathcal{Q}_1 \\ Y_2^2 & \text{given } \mathbf{Y} \in \mathcal{Q}_2 \\ 0 & \text{given } \mathbf{Y} \in \mathcal{Q}_3 \\ Y_1^2 & \text{given } \mathbf{Y} \in \mathcal{Q}_4 \end{cases} \sim \begin{cases} \chi_2^2 \\ \chi_1^2 \\ \chi_0^2 = 0 \\ \chi_1^2 \end{cases}.$$

Then the null distribution of the LRT is the mixture of chi-square distributions for $c > 0$:

$$\begin{aligned} pr(LRT \leq c | H_0) &= 0.25pr(\chi_0^2 \leq c) + 0.5pr(\chi_1^2 \leq c) + 0.25pr(\chi_2^2 \leq c) \\ &= \sum_{i=0}^2 w_i pr(\chi_i^2 \leq c), \end{aligned}$$

where $(w_0, w_1, w_2) = (0.25, 0.5, 0.25)$, which are the probabilities that $\mathbf{Y}$ falls in the cones $\mathcal{Q}_3, \mathcal{Q}_2 \cup \mathcal{Q}_4$, and $\mathcal{Q}_1$, respectively.

We have covered the simple case where the variance-covariance matrix is the identity matrix. We now consider a case for the general positive definite variance-covariance matrix $\mathbf{V}$ to address our inference problem.

Let $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \mathbf{V})$ be a $p \times 1$ normal random vector and $\mathcal{C}$ a closed convex cone in $\mathbb{R}^p$. The kernel of the log-likelihood is:

$$-2\ell(\boldsymbol{\theta}) = \|\mathbf{Y} - \boldsymbol{\theta}\|_{\mathbf{V}}^2.$$

Let $\tilde{\boldsymbol{\theta}}$ be the constrained MLE of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \in \mathcal{C}$. Since $-2\ell(\boldsymbol{\theta})$ is equal to the squared distance between $\mathbf{Y}$ and $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$ is the point in $\mathcal{C}$ that is closest to $\mathbf{Y}$, i.e. $\tilde{\boldsymbol{\theta}}$ is the least squares projection of $\mathbf{Y}$ onto $\mathcal{C}$ denoted as:

$$\tilde{\boldsymbol{\theta}} = P_{\mathbf{V}}(\mathbf{Y}|\mathcal{C}). \quad (5.13)$$

We outline the GP method in Section 4.3.2. We can use this method as an approach to determine $\tilde{\boldsymbol{\theta}}$. The advantage of this method is that it is directly applicable even when $\mathcal{C}$ is a translated cone with vertex other than the origin. We now present distributional properties of the likelihood ratio tests under linear inequality constraints. Define the following parameter spaces:

$$\mathcal{C}_0 = \{\mathbf{z} \in \mathbb{R}^p : \mathbf{z} = \mathbf{0}\},$$

$$\mathcal{C} = \{\mathbf{z} \in \mathbb{R}^p\} \text{ a closed convex cone, and}$$

$$\mathcal{C}_2 = \{\mathbf{z} \in \mathbb{R}^p : \text{no restriction on } \mathbf{z}\}.$$

Consider a testing problem for $H_0 : \boldsymbol{\theta} \in \mathcal{C}_0$ against $H_1 - H_0$, where $H_1 : \boldsymbol{\theta} \in \mathcal{C}$. Then, the LRT is given by:

$$\begin{aligned} \bar{\chi}_{01}^2(\mathbf{V}, \mathcal{C}) &= 2 [\max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}\} - \max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}_0\}], \\ &= 2 [\ell(\tilde{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)] = \|\mathbf{Y} - \boldsymbol{\theta}_0\|_{\mathbf{V}}^2 - \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2 = \|\mathbf{Y}\|_{\mathbf{V}}^2 - \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2, \\ &= \mathbf{Y}\mathbf{V}^{-1}\mathbf{Y} - \min_{\boldsymbol{\theta} \in \mathcal{C}} (\mathbf{Y} - \boldsymbol{\theta})^T \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\theta}) = \mathbf{Y}^T \mathbf{V}^{-1} \mathbf{Y} - (\mathbf{Y} - \tilde{\boldsymbol{\theta}})^T \mathbf{V}^{-1}(\mathbf{Y} - \tilde{\boldsymbol{\theta}}), \\ &= \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2 = \|P(\mathbf{Y}|\mathcal{C})\|_{\mathbf{V}}^2. \end{aligned}$$

The last step follows since $(\mathbf{Y} - \tilde{\boldsymbol{\theta}})^T \mathbf{V}^{-1} \tilde{\boldsymbol{\theta}} = 0$, i.e. $(\mathbf{Y} - \tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}$ are $\mathbf{V}$ -orthogonal. When testing $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 - H_1$, where $H_2 : \boldsymbol{\theta} \in \mathcal{C}_2$, the LRT test statistic is given by:

$$\begin{aligned}
\bar{\chi}_{12}^2(\mathbf{V}, \mathcal{C}) &= 2 [\max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}_2\} - \max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}\}] = 2 [\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})], \\
&= \min_{\boldsymbol{\theta} \in \mathcal{C}} \|\mathbf{Y} - \boldsymbol{\theta}\|_{\mathbf{V}}^2 - \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \|\mathbf{Y} - \boldsymbol{\theta}\|_{\mathbf{V}}^2 = \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2 - \|\mathbf{Y} - \hat{\boldsymbol{\theta}}\|_{\mathbf{V}}^2, \\
&= \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2 - \|\mathbf{Y} - \mathbf{Y}\|_{\mathbf{V}}^2, \\
&= \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2.
\end{aligned}$$

Note that $(\mathbf{Y} - \tilde{\boldsymbol{\theta}})$ is also the point in $\mathcal{C}^\circ$ that is closest to $\mathbf{Y}$, where $\mathcal{C}^\circ$ is the polar cone of $\mathcal{C}$. Since $\mathcal{C}^\circ$ is a closed convex cone in $\mathbb{R}^p$, $\|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2$ is the LRT for testing

$$H_0^* : \boldsymbol{\theta} = \mathbf{0} \text{ against } H_1^* : \boldsymbol{\theta} \in \mathcal{C}^\circ.$$

We have

$$\begin{aligned}
\bar{\chi}_{12}^2(\mathbf{V}, \mathcal{C}) &= \min_{\boldsymbol{\theta} \in \mathcal{C}} (\mathbf{Y} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\theta}) = \|P(\mathbf{Y}|\mathcal{C}^\circ)\|_{\mathbf{V}}^2, \\
&= \mathbf{Y}^T \mathbf{V}^{-1} \mathbf{Y} - \min_{\boldsymbol{\theta} \in \mathcal{C}^\circ} (\mathbf{Y} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\theta}), \\
&= \|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2 = \bar{\chi}^2(\mathbf{V}, \mathcal{C}^\circ).
\end{aligned}$$

Thus, $\|\mathbf{Y} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{V}}^2$ is the LRT for testing $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 - H_1$, where $H_2 : \boldsymbol{\theta} \in \mathcal{C}_2$. Therefore, the null distribution of the LRT for testing $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \in \mathcal{C}_2$, and the null distribution of the LRT for testing $H_0^* : \boldsymbol{\theta} = \mathbf{0}$ against $H_1^* : \boldsymbol{\theta} \in \mathcal{C}^\circ$ are the same. The above proof $\bar{\chi}_{12}^2(\mathbf{V}, \mathcal{C})$ follow from Proposition (5.1) (see [15]) for further details and proof).

Proposition 5.1: Let $\mathcal{C}$ be a closed convex cone and $\mathbf{x} \in \mathbb{R}^p$.

- (1) Assume that $\mathbf{x} = \mathbf{y} + \mathbf{z}$ with $\mathbf{y} \in \mathcal{C}$, $\mathbf{z} \in \mathcal{C}^\circ$ and $\mathbf{y}^T \mathbf{z} = 0$. Then $\mathbf{y} = P(\mathbf{x}|\mathcal{C})$ and $\mathbf{z} = P(\mathbf{x}|\mathcal{C}^\circ)$
- (2) Conversely, $\mathbf{x} = P(\mathbf{x}|\mathcal{C}) + P(\mathbf{x}|\mathcal{C}^\circ)$ and $P(\mathbf{x}|\mathcal{C})^T P(\mathbf{x}|\mathcal{C}^\circ) = 0$.
- (3) $\mathbb{R}^p = \mathcal{C} \oplus \mathcal{C}^\circ$ and $\mathcal{C}^{\circ\circ} = \mathcal{C}$.

Geometry of MLE and LRT when $\mathbf{Y} \sim N(\boldsymbol{\theta}, \mathbf{V})$

Let $\mathcal{C}$ be the convex cone AOB in Figure 5.4 and suppose that $\boldsymbol{\theta}$ is constrained to lie in it. The left plot shows the MLE when $\boldsymbol{\theta}$ is restricted to $\mathcal{C}$ and the right plot shows the critical region for testing H_0 against $H_1 - H_0$. Let OC and OA be $\mathbf{V}$ -orthogonal and OD and OB be $\mathbf{V}$ -orthogonal. Then, COD is the polar cone of $\mathcal{C}$ w.r.t $\langle, \rangle_{\mathbf{V}}$. Let Q and R be the points of intersection of an arbitrary contour ($\mathbf{Y}^T \mathbf{V}^{-1} \mathbf{Y} = \text{constant}$) with OA and OB , respectively. Let PQ and SR be the tangents to the contour at Q and R , respectively. Thus, $PQRS$ is a smooth curve with continuous slope everywhere. If $\mathbf{Y} \in AOB$ then $\tilde{\boldsymbol{\theta}} = \mathbf{Y}$; if $\mathbf{Y} \in COA$, say $\mathbf{Y} = OP$, then $\tilde{\boldsymbol{\theta}} = OQ$; if $\mathbf{Y} \in DOC$ then $\tilde{\boldsymbol{\theta}} = \mathbf{0}$; and if $\mathbf{Y} \in DOB$, say $\mathbf{Y} = OS$, then $\tilde{\boldsymbol{\theta}} = OR$. Now, the boundary of a typical critical region is $PQRS$, where the QR is the segment of the contour $\mathcal{C}$, i.e. ($\mathbf{Y}^T \mathbf{V}^{-1} \mathbf{Y} = \text{constant}$) that lies in AOB .

Figure 5.4: The constrained MLE of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \in \mathcal{C}$ and the LRT of H_0 vs H_1 and a typical boundary of the critical region is $PQRS$

Hence, the chi-bar-squared test statistics $\bar{\chi}_{01}^2(\mathbf{V}, \mathcal{C})$ and $\bar{\chi}_{12}^2(\mathbf{V}, \mathcal{C})$ are expressed in terms of the distance between the origin of $\mathbf{Y}$ and its projection onto a closed convex cone. Many distributional results concerning these models may be stated under the assumption of normality.

These results are well summarized in Section 5.4 below (see [15] for further details).

5.4 CHI-BAR-Square Distribution

The null distributions of likelihood ratio statistics for the above test problems with multivariate normal data turn out to be chi-bar-square. In this section, we introduce the general form of chi-bar-square distribution.

Let $\mathcal{C} \subset \mathbb{R}^p$ (i.e. $\mathcal{C}$ is a closed convex cone) and let $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{V})$, where $\mathbf{V}$ is a positive definite matrix. We define $\bar{\chi}^2(\mathbf{V}, \mathcal{C})$ to be the random variable, which has the same distribution as $\mathbf{Z}\mathbf{V}^{-1}\mathbf{Z} - \min_{\boldsymbol{\theta} \in \mathcal{C}} (\mathbf{Z} - \boldsymbol{\theta})^T \mathbf{V}^{-1}(\mathbf{Z} - \boldsymbol{\theta})$. We also write

$$\bar{\chi}^2(\mathbf{V}, \mathcal{C}) = \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} - \min_{\boldsymbol{\theta} \in \mathcal{C}} (\mathbf{Z} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Z} - \boldsymbol{\theta}). \quad (5.14)$$

Geometric interpretation of $\bar{\chi}^2(\mathbf{V}, \mathcal{C})$

Consider the value $\mathbf{Z}$ represented by OA . Let B be the point in $\mathcal{C}$ that is closest to A , and let $\tilde{\mathbf{Z}}$ denote the vector OB . Therefore,

$$\tilde{\mathbf{Z}} = \min_{\mathbf{x} \in \mathcal{C}} (\mathbf{Z} - \mathbf{x})^T \mathbf{V}^{-1} (\mathbf{Z} - \mathbf{x}) = P_{\mathbf{V}}(\mathbf{Z} | \mathcal{C}).$$

In other words, $\tilde{\mathbf{Z}}$ is the $\mathbf{V}$ -projection of $\mathbf{Z}$ onto $\mathcal{C}$; thus $\mathbf{Z} - \tilde{\mathbf{Z}}$ is $\mathbf{V}$ -orthogonal to $\mathbf{Z}$. In other words, OB is $\mathbf{V}$ -orthogonal to AB .

Figure 5.5: OB and OC are the $\mathbf{V}$ -projections of OA onto $\mathcal{C}$ and $\mathcal{C}^\circ$ respectively.

In triangle OAB , we have

$$\|OA\|_{\mathbf{V}}^2 = \|OB\|_{\mathbf{V}}^2 + \|BA\|_{\mathbf{V}}^2$$

$$\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} = \tilde{\mathbf{Z}}^T \mathbf{V}^{-1} \tilde{\mathbf{Z}} + \min_{\mathbf{x} \in \mathcal{C}} (\mathbf{Z} - \mathbf{x})^T \mathbf{V}^{-1} (\mathbf{Z} - \mathbf{x})$$

Therefore,

$$\bar{\chi}^2(\mathbf{V}, \mathcal{C}) = \tilde{\mathbf{Z}}^T \mathbf{V}^{-1} \tilde{\mathbf{Z}} = \|OB\|_{\mathbf{V}}^2$$

and $\|\mathbf{Z} - \mathcal{C}\|_{\mathbf{V}}$ is the $\mathbf{V}$ -distance between the point $\mathbf{Z}$ and the cone set $\mathcal{C}$ defined by

$$\|\mathbf{Z} - \mathcal{C}\|_{\mathbf{V}}^2 = \min_{\mathbf{x} \in \mathcal{C}} (\mathbf{Z} - \mathbf{x})^T \mathbf{V}^{-1} (\mathbf{Z} - \mathbf{x}) = \|BA\|_{\mathbf{V}}^2.$$

The polar cone $\mathcal{C}^\circ$ of $\mathcal{C}$ with respect to the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}} = \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y}$ is defined by

$$\mathcal{C}^\circ = \{\mathbf{x} : \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y} \leq 0 \quad \forall \mathbf{y} \in \mathcal{C}\}.$$

Because $\mathcal{C}$ is a closed convex cone, $\mathcal{C}^\circ$ is also a closed convex cone. Let C be the point in $\mathcal{C}^\circ$ that is closest to A , and let $\mathbf{Z}^\circ$ denote the vector OC . Therefore,

$$\mathbf{Z}^\circ = \min_{\mathbf{x} \in \mathcal{C}^\circ} (\mathbf{Z} - \mathbf{x})^T \mathbf{V}^{-1} (\mathbf{Z} - \mathbf{x}) = P_{\mathbf{V}}(\mathbf{Z} | \mathcal{C}^\circ).$$

In other words, $\mathbf{Z}^\circ$ is the $\mathbf{V}$ -projection of $\mathbf{Z}$ onto $\mathcal{C}^\circ$. In the rectangle $OBAC$,

$$OC = BA, OB = CA, \|OA\|_{\mathbf{V}}^2 = \|OB\|_{\mathbf{V}}^2 + \|BA\|_{\mathbf{V}}^2 = \|OC\|_{\mathbf{V}}^2 + \|CA\|_{\mathbf{V}}^2.$$

Therefore, it is useful to note that

$$\bar{\chi}^2(\mathbf{V}, \mathcal{C}^\circ) = \|OC\|_{\mathbf{V}}^2 = \|\mathbf{Z} - \mathcal{C}\|_{\mathbf{V}}^2.$$

Proposition 5.2: *Let $\mathbf{V}$ be $p \times p$ positive definite matrix. Then*

$$(1) \quad \|\mathbf{Z}\|_{\mathbf{V}}^2 = \|P_{\mathbf{V}}(\mathbf{Z}|\mathcal{C})\|_{\mathbf{V}}^2 + \|\mathbf{Z} - P_{\mathbf{V}}(\mathbf{Z}|\mathcal{C})\|_{\mathbf{V}}^2,$$

$$(2) \quad P_{\mathbf{V}}(\mathbf{Z}|\mathcal{C}^\circ) = \mathbf{Z} - P_{\mathbf{V}}(\mathbf{Z}|\mathcal{C}).$$

$$(3) \quad \text{If } \mathbf{Z} \sim N(\mathbf{0}, \mathbf{V}), \text{ then}$$

$$\|P_{\mathbf{V}}(\mathbf{Z}|\mathcal{C})\|_{\mathbf{V}}^2 \sim \bar{\chi}^2(\mathbf{V}, \mathcal{C}) \quad \text{and} \quad \|\mathbf{Z} - P_{\mathbf{V}}(\mathbf{Z}|\mathcal{C})\|_{\mathbf{V}}^2 \sim \bar{\chi}^2(\mathbf{V}, \mathcal{C}^\circ).$$

The null distribution for the ordered hypothesis was found to be chi-bar-squared, which is a mixture of chi-squared distributions. Constrained LRT were also derived and shown to follow a chi-bar-square distribution.

Theorem 5.2 (LRT distribution when $\mathbf{Y}$ is normal): *Let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^p$ and $\mathbf{V}$ be a $p \times p$ positive definite matrix. Then under the null hypothesis, the distributions of $\bar{\chi}_{01}^2(\mathbf{V}, \mathcal{C})$ and $\bar{\chi}_{12}^2(\mathbf{V}, \mathcal{C})$ when $\mathbf{Y} \sim N_p(\mathbf{0}, \mathbf{V})$ are given by*

$$pr\{\bar{\chi}_{01}^2(\mathbf{V}, \mathcal{C}) \leq c\} = \sum_{i=0}^p w_i(p, \mathbf{V}, \mathcal{C}) pr(\chi_i^2 \leq c), \quad (5.15)$$

$$pr\{\bar{\chi}_{12}^2(\mathbf{V}, \mathcal{C}) \leq c\} = \sum_{i=0}^p w_{p-i}(p, \mathbf{V}, \mathcal{C}) pr(\chi_i^2 \leq c) = pr\{\bar{\chi}^2(\mathbf{V}, \mathcal{C}^\circ) \leq c\}, \quad (5.16)$$

where $w_i(p, \mathbf{V}, \mathcal{C})$ are some nonnegative numbers and $\sum_{i=0}^p w_i(p, \mathbf{V}, \mathcal{C}) = 1$. When $\mathcal{C}$ is replaced by its polar cone $\mathcal{C}^\circ$, the weights appear in the reverse order.

Details about the quantities $w_i(p, \mathbf{V}, \mathcal{C})$ and their computation are discussed in Section 5.4.1. The right-hand side of equation (5.15) is a weighted mean of several tail probabilities of χ^2 -distributions, and hence is known as a chi-bar-square distribution. We shall refer to $w_i(p, \mathbf{V}, \mathcal{C})$ as chi-bar-square weights or simply as weights. Another term used for these weights is level probabilities. It is worth noting that the $\bar{\chi}^2$ -statistic is based on principles of generalized least squares, and it is therefore a reasonable test statistic even if the distribution of $\mathbf{Y}$ is not normal.

What is the LRT if the null parameter space is replaced by linear space?

We can also derive similar results even when the null parameter space, $\{\mathbf{0}\}$, is replaced by a linear space. Let $\mathcal{M}$ be a linear space contained in $\mathcal{C}$. In particular, if the constraints are a linear inequality, i.e. if we are interested in testing

$$H_0 : \boldsymbol{\theta} \in \mathcal{M} \quad \text{against} \quad H_1 : \boldsymbol{\theta} \in \mathcal{C},$$

which is similar to testing (see [15] for further details)

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{\theta} \in \mathcal{M}^\perp \cap \mathcal{C},$$

where $\mathcal{M}^\perp = \{\mathbf{x} : \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y} = 0 \quad \forall \mathbf{y} \in \mathcal{M}\}$ is the orthogonal complement of $\mathcal{M}$ w.r.t. the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}}$, then we have the following results:

Corollary 5.1: *The LRT for the testing, $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ is similar, i.e. the null distribution of the test statistic is the same at every point in the null parameter space, and its null distribution is given by*

$$pr\{LRT \leq c\} = \sum_{i=0}^p w_i(p, \mathbf{V}, \mathcal{C} \cap \mathcal{M}^\perp) pr(\chi_i^2 \leq c). \quad (5.17)$$

Proof. The proof of this corollary is based on results about projections of $\mathbf{Y} \sim N(\boldsymbol{\theta}, \mathbf{V})$ onto convex cones. A least squares statistic for testing H_0 against H_1 is

$$L = \min_{\boldsymbol{\theta} \in \mathcal{M}} \{q(\boldsymbol{\theta})\} - \min_{\boldsymbol{\theta} \in \mathcal{C}} \{q(\boldsymbol{\theta})\}, \quad (5.18)$$

where

$$q(\boldsymbol{\theta}) = (\mathbf{Y} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\theta}) = \|\mathbf{Y} - \boldsymbol{\theta}\|_{\mathbf{V}}^2 = -2\ell(\boldsymbol{\theta}),$$

and $\mathbf{V}$ is a known positive definite matrix. Since $\mathbf{Y} \sim N(\boldsymbol{\theta}, \mathbf{V})$, the LRT for testing H_0 against H_1 is derived from

$$\begin{aligned}
LRT &= \min_{\boldsymbol{\theta} \in \mathcal{M}} \{\|\mathbf{Y} - \boldsymbol{\theta}\|_{\mathbf{V}}^2\} - \min_{\boldsymbol{\theta} \in \mathcal{C}} \{\|\mathbf{Y} - \boldsymbol{\theta}\|_{\mathbf{V}}^2\}, \\
&= \|\mathbf{Y}\|_{\mathbf{V}}^2 - \min_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{M}^\perp} \{\|\mathbf{Y} - \boldsymbol{\theta}\|_{\mathbf{V}}^2\}.
\end{aligned}$$

□

Since $\mathcal{C} \cap \mathcal{M}^\perp$ is a closed convex cone, its null distribution is a chi-bar-square and is given by Eq. (5.17). In general, the weights of the chi-bar-square distribution in Eq. (5.17) depend on the parameter spaces $\mathcal{M}$ and $\mathcal{C}$. There is no easy way to compute these weights for an arbitrary $\{\mathcal{C}, \mathcal{M}, \mathbf{V}\}$. The simulation procedure indicated in Section 5.4.1 is available as a general purpose procedure for computing $\text{pr}\{LRT \leq c | H_0\}$.

Theorem 5.3: Let $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \mathbf{V})$, where $\mathbf{V}$ is a $p \times p$ positive definite matrix, $\mathbf{R}$ be a row-full-rank matrix of order $r \times p$, $\text{rank}(\mathbf{R}) = r \leq p$, and let $\mathbf{R}_1$ be a submatrix of $\mathbf{R}$ of order $q \times p$. Let the hypotheses be $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}$, $H_1 : \mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}$ and $H_2 : \text{no restrictions on } \boldsymbol{\theta}$, respectively. Then, the LRT statistics $\bar{\chi}_{01}^2$ and $\bar{\chi}_{12}^2$ for testing H_0 versus $H_1 - H_0$ and H_1 versus $H_2 - H_1$ respectively, have null distributions under H_0 given by

$$\text{pr}\{\bar{\chi}_{01}^2 \leq c\} = \sum_{i=0}^q w_i(q, \mathbf{R}_1 \mathbf{V} \mathbf{R}_1^T, \mathcal{C}) \text{pr}(\chi_{r-q+i}^2 \leq c), \quad (5.19)$$

$$\text{pr}\{\bar{\chi}_{12}^2 \leq c\} = \sum_{i=0}^q w_{q-i}(q, \mathbf{R}_1 \mathbf{V} \mathbf{R}_1^T, \mathcal{C}) \text{pr}(\chi_i^2 \leq c), \quad (5.20)$$

where $\mathcal{C} = \{\mathbf{z} \in \mathbb{R}^q : z_i \geq 0, i = 1, \dots, q\}$.

If the alternative hypothesis does not have any inequality constraints, then $q = 0$ and hence the classical chi-square test with r degrees of freedom is obtained as $\text{pr}(LRT \leq c | H_0) = \text{pr}(\chi_r^2 \leq c)$.

Lemma 5.1: Note that the number of terms in the above chi-bar-square distribution depends on the number of inequalities in H_1 only, not on the dimension p of $\boldsymbol{\theta}$.

One distinguishing factor of the $\bar{\chi}_{12}^2$ test is that the null hypothesis involves inequalities. Hence the p -value depends on the underlying parameter $\boldsymbol{\theta}$, which may be anywhere in the null parameter space $\{\boldsymbol{\theta} : \mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}\}$, for example. However, in order to obtain the critical value, c , which assures size α , we must solve $\sup_{\mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}} P_{\boldsymbol{\theta}}(\bar{\chi}_{12}^2 > c) = \alpha$. As explained in [15], the supremum occurs at any $\boldsymbol{\theta}_0$ with $\mathbf{R}_1\boldsymbol{\theta}_0 = \mathbf{0}$, and hence $\boldsymbol{\theta} = \mathbf{0}$ is one such case. This particular null distribution is denoted as the least favorable distribution.

If the alternative hypothesis has only independent linear inequalities, and the number of parameters is small and does not exceed 3 (i.e. $p \leq 3$), then we can use the explicit formulas for weights w_i . Such closed-form weight expressions were given by Kudo [68] and Silvapulle [69]. If the number of parameters $p \geq 4$, **simulated** weights may be used, as the LRT p -value is not sensitive to the weights. A standard approach to simulate the chi-bar-square weights $w_i(p, \mathbf{V}, \mathcal{C}), i = 0 \cdots p$, is given in Section 5.4.1 below:

5.4.1 CHI-BAR-SQUARE Weights

As indicated earlier, the null distribution of several test statistics for or against inequality constraints turns out to be $\bar{\chi}^2$. Therefore, we need to be able to compute its tail probability to obtain the p -value and/or critical value. This would be easy if the chi-bar-square weights $\{w_i\}$ are known. The chi-bar-square weights, $w_i(p, \mathbf{V})$, also known as probabilities level, represent the probability that the least squares projection of a p -dimensional multivariate normal observation from $N_p(\mathbf{0}, \mathbf{V})$ onto the positive orthant cone has exactly i positive component values. Unfortunately, the exact computation of $\{w_i\}$ is quite difficult in general. However, we can compute the tail probability of a chi-bar-square distribution by simulations.

Algorithm 5.1 (To compute $pr\{\bar{\chi}^2(\mathbf{V}, \mathcal{C}) \geq c\}$): *The following steps are used to compute the tail probabilities of $\bar{\chi}^2$ -distribution:*

- (1) Generate $\mathbf{Z}$ from $N_p(\mathbf{0}, \mathbf{V})$.
- (2) Compute $\bar{\chi}^2(\mathbf{V}, \mathcal{C})$ in Eq. (5.14).

- (3) Repeat the first two steps N times (say, $N = 10000$).
- (4) Estimate $pr\{\bar{\chi}^2(\mathbf{V}, \mathcal{C}) \geq c\}$ by (n/N) where n is the number of times $\bar{\chi}^2(\mathbf{V}, \mathcal{C})$ in the second step that turned out to be greater than or equal to c .

Algorithm 5.2 (Simulation to compute $w_i(p, \mathbf{V}, \mathcal{C})$ when $\mathcal{C}$ is the polyhedral): The following steps are used to compute the chi-bar-square weights $w_i(p, \mathbf{V}, \mathcal{C}), i = 0 \cdots p$, of $\bar{\chi}^2$ -distribution:

- (1) Generate $\mathbf{Z}$ from $N_p(\mathbf{0}, \mathbf{V})$.
- (2) Compute $\tilde{\mathbf{Z}}$, (see below), the point at which $(\mathbf{Z} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Z} - \boldsymbol{\theta})$ is a minimum over $\boldsymbol{\theta} \in \mathcal{C}$. (For the purpose of this thesis, the “solve.QP” built-in R software function is used).
- (3) Compute s , where s is the dimension of set, $\phi = \{\boldsymbol{\theta} : \mathbf{a}_j^T \boldsymbol{\theta} = 0 \quad \forall j \in J\}$, of the active constraints at the solution, where $J = \{j : \mathbf{a}_j^T \tilde{\mathbf{Z}} = 0\}$ and $\mathbf{a}_j^T s$ define the polyhedral.
- (4) Repeat the first three steps N times (say, $N = 10000$).
- (5) Estimate $w_i(p, \mathbf{V}, \mathcal{C})$ by (n_i/N) where n_i is the number of times s is exactly equal to i , ($i = 0 \cdots p$).

Note that whenever the cone $\mathcal{C}$ is the **positive orthant**, i.e. $\mathcal{C} = \mathbb{R}^{+p}$, we write

$$w_i(p, \mathbf{V}, \mathcal{C}) = w_i(p, \mathbf{V}). \quad (5.21)$$

Algorithm 5.3 (Simulation Algorithm to compute $w_i(p, \mathbf{V}), i = 0 \cdots p$): The following steps are used to compute the chi-bar-square weights $w_i(p, \mathbf{V}), i = 0 \cdots p$, of $\bar{\chi}^2$ -distribution:

- (1) Generate $\mathbf{Z}$ from $N_p(\mathbf{0}, \mathbf{V})$.
- (2) Compute $\tilde{\mathbf{Z}}$, (see below), the point at which $(\mathbf{Z} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Z} - \boldsymbol{\theta})$ is a minimum over $\boldsymbol{\theta} \geq \mathbf{0}$.
- (3) Count the number of positive components of $\tilde{\mathbf{Z}}$. (This is equal to s in Step 3 of Simulation 5.2).

- (4) Repeat the first three steps N times (say, $N = 10000$).
- (5) Estimate $w_i(p, \mathbf{V}, \mathbb{R}^{+p})$ by (n_i/N) where n_i is the number of times $\tilde{\mathbf{Z}}$ has exactly i positive components, $(i = 0 \cdots p)$.

If $\mathcal{C}$ involves only linear constraints, then a quadratic program can be used for computing

$$\tilde{\mathbf{Z}} = \min_{\boldsymbol{\theta} \geq \mathbf{0}} \{(\mathbf{Z} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\mathbf{Z} - \boldsymbol{\theta})\} = \min_{\boldsymbol{\theta} \geq \mathbf{0}} g(\boldsymbol{\theta}).$$

Suppose that we wish to solve

$$\min g(\boldsymbol{\theta}) \quad \text{subject to } \mathbf{A}_1 \boldsymbol{\theta} \geq \mathbf{0} \text{ and } \mathbf{A}_2 \boldsymbol{\theta} = \mathbf{0},$$

for some matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ that do not depend on $\boldsymbol{\theta}$. This constrained minimization problem in which the objective function is quadratic in $\boldsymbol{\theta}$ and the constraints are linear equality and inequality constraints in $\boldsymbol{\theta}$ is called a quadratic program. In this thesis, we used the R software built-in function “solve.QP” for this optimization problem. This quadratic programming problem is sometimes expressed in the following, slightly different but equivalent, form. Note that

$$g(\boldsymbol{\theta}) = 2f(\boldsymbol{\theta}) + \text{constant}, \text{ where } f(\boldsymbol{\theta}) = \mathbf{a}^T \boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{V}^{-1} \boldsymbol{\theta},$$

and $\mathbf{a} = -(\mathbf{V}^{-1})^T \mathbf{Z}$. Therefore, the minimization of $g(\boldsymbol{\theta})$ subject to some constraints on $\boldsymbol{\theta}$ is equivalent to the minimization of $f(\boldsymbol{\theta})$ subject to the same constraints on $\boldsymbol{\theta}$. Therefore, $\tilde{\mathbf{Z}}$, the solution to $\min g(\boldsymbol{\theta})$, is also the solution to $\min f(\boldsymbol{\theta})$ subject to $\mathbf{A}_1 \boldsymbol{\theta} \geq \mathbf{0}$ and $\mathbf{A}_2 \boldsymbol{\theta} = \mathbf{0}$. To access R code to compute the chi-bar-square weights see Appendix F.

The following theorem provides some theoretical results concerning the chi-bar-square weights $w_i(p, \mathbf{V}, \mathcal{C})$ that may be applied when computing or simulating values.

Theorem 5.4: *Let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^p$ and $\mathbf{V}$ be a $p \times p$ nonsingular covariance matrix. Then we have the following:*

- (1) Let $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{V})$ and $\mathcal{C}$ be the nonnegative orthant. Then
 $w_i(p, \mathbf{V}, \mathcal{C}) = \text{pr}\{P_{\mathbf{V}}(\mathbf{Z}|\mathcal{C}) \text{ has exactly } i \text{ positive components}\},$
- (2) $\sum_{i=0}^p (-1)^i w_i(p, \mathbf{V}, \mathcal{C}) = 0,$
- (3) $0 \leq w_i(p, \mathbf{V}, \mathcal{C}) \leq 0.5.$
- (4) Let $\mathcal{C}^\circ$ denote the polar cone, $\{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}\mathbf{V}^{-1}\mathbf{y} \leq 0 \quad \forall \mathbf{y} \in \mathcal{C}\}$, of $\mathcal{C}$ w.r.t the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}} = \mathbf{x}^T \mathbf{V}^T \mathbf{y}$. Then $w_i(p, \mathbf{V}, \mathcal{C}^\circ) = w_{p-i}(p, \mathbf{V}, \mathcal{C}).$
- (5) Let $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}\}$, where $\mathbf{R}$ is a $k \times p$ of rank $k(\leq p)$. Then, $\bar{\chi}^2(\mathbf{V}, \mathcal{C}) = \bar{\chi}^2(\mathbf{R}\mathbf{V}\mathbf{R}^T, \mathbb{R}^{+p})$ and $w_i(p, \mathbf{V}, \mathcal{C}) = w_i(p, \mathbf{R}\mathbf{V}\mathbf{R}^T).$
- (6) $w_i(p, \mathbf{V}) = w_{p-i}(p, \mathbf{V}^{-1}).$
- (7) Let $\mathbf{C}$ be the correlation matrix corresponding to $\mathbf{V}$. Then, $\bar{\chi}^2(\mathbf{V}, \mathcal{C}) = \bar{\chi}^2(\mathbf{C}, \mathcal{C})$ and $w_i(p, \mathbf{V}, \mathcal{C}) = w_i(p, \mathbf{C}, \mathcal{C})$ for every i .
- (8) Let $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}\}$, where $\mathbf{R}$ is a $q \times p$ matrix of row full rank $q(\leq p)$. Then

$$w_{p-q+i}(p, \mathbf{V}, \mathcal{C}) = \begin{cases} w_i(q, \mathbf{R}\mathbf{V}\mathbf{R}^T) & \text{for } i = 0, \dots, q \\ 0 & \text{otherwise} \end{cases}$$

- (9) Let $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}, \mathbf{R}_2\boldsymbol{\theta} = \mathbf{0}\}$, where $\mathbf{R}_1$ is a $s \times p$, $\mathbf{R}_2$ is a $t \times p$, $s+t \leq p$, $[\mathbf{R}_1^T, \mathbf{R}_2^T]$ is of full row rank matrix, and

$$\mathbf{V}_{new} = \mathbf{R}_1\mathbf{V}\mathbf{R}_1^T - (\mathbf{R}_1\mathbf{V}\mathbf{R}_2^T)(\mathbf{R}_2\mathbf{V}\mathbf{R}_2^T)^{-1}(\mathbf{R}_2\mathbf{V}\mathbf{R}_1^T).$$

Then

$$w_{p-s-t+i}(p, \mathbf{V}, \mathcal{C}) = \begin{cases} w_i(s, \mathbf{V}_{new}) & \text{for } i = 0, \dots, s \\ 0 & \text{otherwise} \end{cases}$$

(10) Let $\mathbf{R}$ be a $r \times p$ matrix of full row rank r , $\mathbf{R}_1$ be a $q \times p$ submatrix of $\mathbf{R}$, $\mathcal{M} = \{\boldsymbol{\theta} : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}\}$, $\mathcal{C} = \{\boldsymbol{\theta} : \mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}\}$, and $\mathcal{M}^\perp$ be the orthogonal complement of $\mathcal{M}$ w.r.t $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}}$. Then

$$w_{r-q+i}(p, \mathbf{V}, \mathcal{C} \cap \mathcal{M}^\perp) = \begin{cases} w_i(q, \mathbf{R}_1 \mathbf{V} \mathbf{R}_1^T) & \text{for } i = 0, \dots, q \\ 0 & \text{otherwise} \end{cases}$$

In the next chapter, we discuss the GP algorithm for inference with categorical data under linear inequality constraints.

Chapter 6

Inference for Categorical Data Under Linear Inequality Constraints

In various fields, such as epidemiology, economics, medicine, etc., it is common for data to have a non-normal distribution. Given the nature of observations being not normally distributed, use of GLM and MGLM models is needed to handle this type of non-normal categorical data. This chapter extends the constrained ML estimation and LR tests for normal data in Chapter 5 to constrained inference in categorical data for GLM (binary data) and MGLM (multinomial data).

The GP algorithm is used to obtain the constrained MLE for the binary and multinomial data (see Sections 6.2 and 6.5). The asymptotic distribution for the constrained LRT is derived, which follows a chi-bar-square distribution (a.k.a. weighted chi-square), see Theorem 6.1. The work leading to the constrained MLE and LRT of the multinomial logit model was progressively achieved by using the techniques and the asymptotic distribution for the binary GLM with minor modifications to the GP algorithm (see Section 6.4).

In this chapter, we detail the computations of constrained ML estimators and hypothesis tests in the generalized logistic regression for binary and multinomial response variables. For the purpose of this thesis, we only work with the nominal responses called baseline-category logit models. The results obtained in this chapter demonstrate the success of our model and the effectiveness of the MLE and LRT under constraints. For information about the derivation of the unrestricted MLEs and the likelihood, we refer to Chapter 3 and the simulation results presented in Section 3.5.

6.1 Generalized Linear Model

The generalized linear model (GLM) is the extension of the linear model (LM) when the observations are discrete or categorical [42]. The LM has the following three assumptions: (i) the observations are independent, (ii) the mean of the observations is a linear function of some covariates, and (iii) the variance of the response variable is a constant. The extension to GLM consists of modifying (ii) and (iii) above; by (ii_a) the mean of the observation is associated with a linear function of some covariates through a link function; and (iii_a) the variance of the observation is a function of the mean. Note that (iii_a) is a result of (ii_a). See McCullagh and Nelder (1989) [42] for details.

The GLM is used to unify various statistical models in order to yield more favourable results. Unlike linear models, GLMs include a variety of models that include normal, binomial, Poisson, and multinomial distributions as special cases. When a GLM is used, the distribution of the response variable ($\mathbf{Y}$) must belong to an exponential family (EF). For more details on the EF and its properties, see Appendix B.

The GLM extends the LM by $\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ and $Y_i \sim EF(\mu_i, \phi)$, where ϕ is a scale (dispersion) parameter and g is a monotone link function. The GLM is related to the expected value of the response $E(Y_i) = \mu_i$ to a linear prediction η_i via the link function $g(\cdot)$.

Let $g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) = \mu_i$ be the inverse link function. For any random variable $Y|X$ with a pdf $f_Y(y; g^{-1}(\mathbf{x}^T \boldsymbol{\beta}), \phi)$, which depends on a canonical parameter $\mu = g^{-1}(\mathbf{x}^T \boldsymbol{\beta})$, the probability density function (pdf) is then considered a member of the EF if:

$$f_Y(y; \mu, \phi) = \exp \left\{ \frac{s(y)\mu - a(\mu)}{b(\phi)} + c(y, \phi) \right\}, \quad (6.1)$$

for some functions a , b and c .

Using the nice property of the EF distribution, we can easily obtain the mean and the variance for the random variable Y . To obtain the mean and variance, we use the log-likelihood of $f_Y(y; \mu, \phi)$ denoted below:

$$\ell(\mu, \phi, y) = \log(f_Y(y; \mu, \phi)) = \frac{s(y)\mu - a(\mu)}{b(\phi)} + c(y, \phi). \quad (6.2)$$

We know that $E\left(\frac{\partial \ell}{\partial \mu}\right) = 0$, and hence the expected value of Y is obtained as

$$E(Y) = \mu = a'(\mu) = a'(g^{-1}(\mathbf{x}^T \boldsymbol{\beta})). \quad (6.3)$$

We also know that $E\left(\frac{\partial^2 \ell}{\partial \mu^2}\right) = -E\left(\frac{\partial \ell}{\partial \mu}\right)^2$, and hence the variance of Y is obtained as

$$V(Y) = a''(\mu)b(\phi) = a''(g^{-1}(\mathbf{x}^T \boldsymbol{\beta}))b(\phi). \quad (6.4)$$

Generalized linear models invoke a mean-variance relationship as a consequence of the link function. The link function in a GLM relates the expected value Y to the covariates. Each member of the EF has a different canonical link function as a result of its specific distribution. In Table 6.1, we list some of these distributions, their link functions as well as the mean functions.

Table 6.1: Exponential Family of Distributions

Exponential Distribution	Link Name	Link Function	Mean Function
Bernoulli	Logit	$\mathbf{x}^T \boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+\exp(-\mathbf{x}^T \boldsymbol{\beta})}$
Binomial	Logit	$\mathbf{x}^T \boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+\exp(-\mathbf{x}^T \boldsymbol{\beta})}$
Exponential	Inverse	$\mathbf{x}^T \boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{x}^T \boldsymbol{\beta})^{-1}$
Gamma	Inverse	$\mathbf{x}^T \boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{x}^T \boldsymbol{\beta})^{-1}$
Inverse Gaussian	Inverse squared	$\mathbf{x}^T \boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{x}^T \boldsymbol{\beta})^{-\frac{1}{2}}$
Normal	Identity	$\mathbf{x}^T \boldsymbol{\beta} = \mu$	$\mu = \mathbf{x}^T \boldsymbol{\beta}$
Multinomial	Logit	$\mathbf{x}^T \boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+\exp(-\mathbf{x}^T \boldsymbol{\beta})}$
Poisson	Log	$\mathbf{x}^T \boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{x}^T \boldsymbol{\beta})$

In GLM, the unknown parameter p -vector $\boldsymbol{\beta}$ is commonly estimated using the method of unrestricted MLE. Here we will fit a **Bernoulli** GLM with logit link function to predict a set of binary observed values 1, 0, i.e. if a patient has heart disease or not, respectively. We fit the GLM to predict the response variable by replacing the expected value of $\mathbf{Y}$ with $\hat{\boldsymbol{\mu}}$.

6.1.1 Unrestricted Inference in GLM

As shown in Chapter 3 the Newton-Raphson algorithm may be used to find the MLE for the coefficient β in the GLM. In general, there are two popular methods to find the MLE in GLM:

- (1) the NR algorithm that maximizes the likelihood function directly, or
- (2) the quasi-likelihood method that specifies only the mean and variance relationship, rather than the full likelihood.

Suppose $\beta^T = (\beta_1^T, \beta_2^T)$ and interest lies in β_1 . Suppose the null hypothesis is given as

$$H_0 : \beta_1 = \beta_{10}.$$

For hypotheses test, the *residual deviance* function is defined as twice the unconstrained log-likelihood ratio, and is given by

$$D(c, s) = -2 \log \left(\frac{L_c(\beta_{10}, \hat{\beta}_{20})}{L_s(\hat{\beta})} \right), \quad (6.5)$$

where

- * $\hat{\beta}_{20}$ is the MLE of β_2 under the restriction that $\beta_1 = \beta_{10}$,
- * $\hat{\beta}^T = (\hat{\beta}_1^T, \hat{\beta}_2^T)$ is the MLE of β^T , and
- * L_c and L_s are the likelihood for the current model and the saturated model, respectively.
- * The random variable $D(c, s)$ is asymptotically distributed as χ_{n-p}^2 , where
 - p is the number of fitted parameters, and
 - n is the sample size.

6.2 Restricted Estimation for Binary Data Using GP

Silvapulle and Sen fully develop the problem of constrained inference for GLMs [15]. Moreover, in Davis' work [10], various authors are referenced in regards to this problem. She also references Bayesian inference for GLMs under a simple order constraint, $\{\boldsymbol{\beta} : \beta_1 < \beta_2 < \dots < \beta_p\}$, where the focus is on deriving a posterior distribution for the means under the order restriction, and performing inferences. Jamshidian (2004) builds on previous works to improve the methods of restricted estimation based on the general likelihood subject to equality $\mathbf{A}_1\boldsymbol{\beta} = \mathbf{b}_1$ and inequality constraints $\mathbf{A}_2\boldsymbol{\beta} \leq \mathbf{b}_1$.

In this section, we use the log-likelihood and its gradient for the binary data derived in Section 3.2. The log-likelihood is obtained as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^n \log(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{n}^T \log\{\mathbf{1} + \exp(\mathbf{X}\boldsymbol{\beta})\}.$$

The score equation is obtained as

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) = 0$$

and the variance-covariance matrix is obtained as the inverse of the Fisher information in Eq. (3.2)

$$\mathbf{W} = (\mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}) \mathbf{X})^{-1}.$$

Remark 6.1 (Constrained vs. Unconstrained MLEs): By definition 4.4, we know that the set of active constraints form a convex cone. If the unconstrained MLE $\hat{\boldsymbol{\beta}}_r$ is inside the cone, i.e. if it satisfies the constraints $\mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}$, so that $\hat{\boldsymbol{\beta}}_r \in \Omega$, then the constrained and unconstrained MLEs are the same.

For the purpose of this work, I computed the unrestricted MLEs using the built-in R function “GLM” with the family link “logit”. To find the restricted MLEs, I adopt the steps below.

- (1) If the unrestricted MLE $\hat{\boldsymbol{\beta}}_r$ satisfies the constraints, then I set the restricted MLE $\tilde{\boldsymbol{\beta}}_r$ as

unrestricted MLE (i.e. $\tilde{\beta}_r = \hat{\beta}_r$) and end the process.

- (2) If the unrestricted MLE is not feasible, an initial working set $\mathbb{W}$ is formed based on either **Scenario i** or **Scenario ii** shown below, along with the corresponding coefficient matrix $\mathbf{A}$ and vector $\mathbf{b}$. Then proceed with the GP algorithm starting at the initial value $\beta_0 \in \Omega$ to find the restricted MLE.

Scenario i) None of the constraints are satisfied by the unrestricted MLE. In this case, choose all active constraints individually to form the working set $\mathbb{W}$, and proceed with the GP method; if one of the active constraints yields a feasible restricted MLE, end the process. If not, then choose all combinations of two constraints as active to form $\mathbb{W}$; if one of these combinations yields a feasible restricted MLE, end the process. Continue this cycle with a combination of three, and so on, if the constraints exist.

Scenario ii) Some of the constraints are satisfied by the unrestricted MLE. We start with choosing the constraints that are not satisfied by the unrestricted MLE. Similarly to **Scenario i**, we proceed with choosing those constraints as active individually, then all combinations of two, then three, and so on, if they exist. If this does not yield a feasible restricted MLE, we proceed with choosing constraints that are satisfied by the unrestricted MLE individually, then all combinations of two, then three and so on, if they exist. If this does not yield a feasible restricted MLE, we continue the process by choosing all constraints (satisfied or unsatisfied) in combinations of two or more, if they exist.

NOTE - if no feasible restricted MLE is found after using all constraints (this is highly unlikely), then we should recheck the convergence algorithm for various reasons, the most common of which is floating points and the initial starting point β_0 .

6.2.1 Empirical Results for Constrained MLE for Binary Data

We ran a series of simulations using the binary regression model to assess the performance of the GP algorithm. For this analysis, the GP algorithm was implemented using the statistical software, R [24]. The study was conducted by generating 1000 datasets, using two different sample sizes of $n = 100$ and $n = 300$ for each set of parameters described below. For Bernoulli(μ_i) regression model:

- The canonical parameter μ_i is modeled as $\mu_i = \pi_i = \frac{\exp \eta_i}{1 + \exp \eta_i}$, with $\eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}$ being the linear predictor.
- The explanatory variables $\mathbf{x}_i = (x_{1i}, x_{2i})^T$ were generated from a bivariate normal distribution with mean $\boldsymbol{\mu} = (2, 1)^T$ and variance-covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}$.

Two constraints on the parameters are imposed as follows:

$$\begin{aligned} \beta_0 + \beta_1 + \beta_2 &\leq 0 \\ -\beta_0 + \beta_1 - \beta_2 &\leq 5 \end{aligned} \Rightarrow \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 5 \end{pmatrix} \quad (6.6)$$

Using both restricted and unrestricted methods, we compare the simulated sampled mean, bias and mean square error (MSE) of the estimates of parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$. We examine several cases of preliminary points $\boldsymbol{\beta}$, where:

- (a) both constraints are active $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ (i.e. both hold with equality),
- (b) at least one constraint is inactive $\mathbf{a}_i\boldsymbol{\beta} < b_i$ (i.e. at least one is a strict inequality), and
 - $\mathbf{b}_1$: The first constraint/prior-knowledge in Eq. (6.6) is inactive.
 - $\mathbf{b}_2$: The second constraint/prior-knowledge in Eq. (6.6) is inactive.
- (c) both constraints are inactive $\mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}$ (i.e. both are within the constraint cone).

Based on our existing knowledge of normal models with order-constrained inference [15] and [10], we would expect the restricted estimators to have a larger bias and a smaller MSE,

particularly when the values of β are near the boundary. In these instances, the unrestricted estimators will likely be found outside the cone. Since the estimates are always forced toward the cone by projecting the estimators onto the closed convex cone, we end up with a bias. The results for unrestricted and restricted estimates are shown in Tables 6.2 and 6.3 for sample sizes of 100 and 300 observations, respectively.

Table 6.2: Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for Bernoulli Model ($n = 100$)

Case	Parameter	True Value	Unrestricted			Restricted		
			Mean	Bias	MSE	Mean	Bias	MSE
a	β_0	-2.00	-2.2651	-0.2651	1.2884	-1.9758	0.0242	0.1556
	β_1	2.50	2.7510	0.2510	0.6561	2.3643	-0.1358	0.0516
	β_2	-0.50	-0.5358	-0.0358	0.1987	-0.4617	0.0384	0.1272
b ₁	β_0	-2.50	-2.6364	-0.1364	0.7750	-2.2320	0.2680	0.2903
	β_1	1.25	1.3230	0.0730	0.1610	1.1347	-0.1153	0.0506
	β_2	-1.25	-1.3252	-0.0752	0.1320	-1.2496	0.0004	0.0934
b ₂	β_0	-2.00	-2.1338	-0.1338	0.5908	-2.3416	-0.3416	0.4974
	β_1	1.00	1.0571	0.0571	0.1172	1.1388	0.1388	0.1027
	β_2	1.00	1.0696	0.0696	0.1141	1.0540	0.0540	0.1074
c	β_0	-2.00	-2.1460	-0.1460	0.5923	-2.2359	-0.2359	0.4918
	β_1	0.90	0.9572	0.0572	0.1123	0.9918	0.0918	0.0967
	β_2	0.90	0.9607	0.0607	0.0972	0.9549	0.0549	0.0949

Table 6.3: Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for Bernoulli Model ($n = 300$)

Case	Parameter	True Value	Unrestricted			Restricted		
			Mean	Bias	MSE	Mean	Bias	MSE
a	β_0	-2.00	-2.0664	-0.0664	0.2511	-1.9640	0.0360	0.0487
	β_1	2.50	2.5623	0.0623	0.1467	2.4096	-0.0904	0.0230
	β_2	-0.50	-0.5101	-0.0101	0.0492	-0.4903	0.0097	0.0364
b ₁	β_0	-2.50	-2.5496	-0.0496	0.1795	-2.3672	0.1328	0.0813
	β_1	1.25	1.2722	0.0222	0.0401	1.1805	-0.0696	0.0157
	β_2	-1.25	-1.2743	-0.0243	0.0445	-1.2209	0.0291	0.0318
b ₂	β_0	-2.00	-2.0489	-0.0489	0.1733	-2.1635	-0.1635	0.1372
	β_1	1.00	1.0269	0.0269	0.0359	1.0736	0.0736	0.0304
	β_2	1.00	1.0181	0.0181	0.0358	1.0034	0.0034	0.0349
c	β_0	-2.00	-2.0297	-0.0297	0.1462	-2.0549	-0.0549	0.1256
	β_1	0.90	0.9180	0.0180	0.0280	0.9281	0.0281	0.0251
	β_2	0.90	0.9160	0.0160	0.0295	0.9130	0.0130	0.0293

As expected, we notice from Tables 6.2 and 6.3 that in most cases, the restricted MLEs have a larger bias and smaller MSE than their unrestricted counterparts; this is especially true for cases b₂ and c. Additionally, with an increasing sample size from 100 to 300 observations, the asymptotic bias and MSE significantly decreased for both restricted and unrestricted MLEs. We note that for restricted estimates, more so than for unrestricted estimates, when the true parameter β is chosen closer to the boundary points of the convex cone, the bias increases and the MSE decreases.

The usual unrestricted MLE is not necessarily suitable when the generating/preliminary point β is not within the cone as it does not incorporate prior knowledge given by (6.6). This means that we cannot determine a targeted direction for the parameters [27]. Table 6.4 shows the percentage of the unrestricted estimates that lie within the cone (i.e. satisfy the constraints).

According to Remark 6.1 on page 103, the restricted MLE is therefore considered the same as the unrestricted MLE in such cases.

Table 6.4: Percentage of unrestricted MLE that satisfy the constraints

Cases	a	b ₁	b ₂	c
$n = 100$	15%	46%	49%	72.5%
$n = 300$	16%	48%	50%	84%

6.3 Constrained Tests for GLM

We extend the unrestricted hypothesis testing for the GLM to the case of linear equality and inequality constraints on the regression parameters. If $\mathbf{Y}$ is not normal, then the statistic L in (5.18) can be used for testing $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$; and would now be defined as a **generalized** least squares based statistic. If H_0 is true, then the distribution of L is the same at every value in the null parameter space, $\mathcal{M}$.

Lemma 6.1: For a large sample $\mathbf{X}_1, \dots, \mathbf{X}_n$, the null distribution for the statistic L is approximately chi-bar-square $\bar{\chi}^2$.

To show that, suppose the sample is independently and identically distributed as $\mathbf{X}$, with $E(\mathbf{X}) = \mathbf{0}$ and $V(\mathbf{X}) = \mathbf{W}$; the distribution of $\mathbf{X}$ may not be normal. Then, for large n , and for $\mathbf{Y} = \frac{1}{n} \sum \mathbf{X}_i$, by multivariate Central Limit Theorem (CLT), $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \mathbf{V} = n^{-1}\mathbf{W})$. Therefore, the statistic L in (5.18) can be used for testing $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, even if it is not the likelihood ratio statistic.

The constrained or restricted tests are associated with the hypotheses:

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b}, \quad H_1 : \mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}, \quad H_2 : \text{no restriction on } \boldsymbol{\beta} \quad (6.7)$$

The restricted ML estimators (RMLE) are obtained by two possible methods: the GP algorithm or the iteratively reweighted-least squares-quadratic programming (IRWLS-QP) approach. There are advantages to both approaches. The GP algorithm is detailed in Sections 4.3.2 and 5.3.2 on pages 55 and 80, respectively. The advantage of the GP algorithm is that it is directly applicable even when $\mathcal{C}$ is a translated cone with vertex other than the origin. Moreover, the GP algorithm can be applied to almost any MLE problem that requires incorporation of equality and inequality constraints [11] and [27]. When the convex cone $\mathcal{C}$ involves only linear equality and inequality constraints, we can use the IRWLS-QP approach in the GLM. When modeling multinomial data, caution must also be taken since the IRWLS-QP approach may be too complicated to implement and/or be inefficient in the time it takes to converge when compared to the GP algorithm. Details on the IRWLS-QP approach are outlined below.

Algorithm 6.1 (IRWLS-QP): *The following steps are used to find the RMLE for GLM using IRWLS-QP. The procedure necessitates an initial guess of $\boldsymbol{\beta}$ called $\boldsymbol{\beta}^0$, and hence $\boldsymbol{\pi}^0$, the probability vector for binary data.*

- (1) *Using the above initial guess, compute an adjusted response variable*

$$z_i = \hat{\eta}_i + (y_i - \hat{\pi}_i) \frac{d\eta_i}{d\pi_i} \Big|_{\hat{\eta}_i} = \text{logit}(\hat{\pi}_i) + \frac{(y_i - n_i \hat{\pi}_i)}{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}.$$

- (2) *Form the weights $1/w_i^0 = \left(\frac{d\eta_i}{d\pi_i} \right)^2 \Big|_{\hat{\eta}_i} V(\hat{\pi}_i^0) \Rightarrow w_i = n_i \hat{\pi}_i (1 - \hat{\pi}_i)$. We use the z_i and w_i as an input to the quadratic programming function as follows:*

- (i) *Compute the least square estimate vector $\hat{\boldsymbol{\beta}}_l$ for $\boldsymbol{\beta}$, and*
- (ii) *compute the $p \times p$ positive definite matrix $\mathbf{D} = \mathbf{X}^T \mathbf{X}$, where $\mathbf{X}$ is the $n \times p$ covariate matrix. The unscaled covariance matrix $\mathbf{D}$ is used because the solution is invariant of covariance matrix.*
- (iii) *Compute the vector $\mathbf{d} = \hat{\boldsymbol{\beta}}_l^T \mathbf{D}$.*
- (iv) *Iterate the built-in R function “solve.QP” to update $\hat{\boldsymbol{\beta}}_l^T$, which in turn updates $\mathbf{d}$.*

(v) Continue the last step until convergence.

(3) Reestimate to get β , and hence π .

Repeat these steps until convergence.

For the purposes of this thesis, the RMLEs were obtained through the GP algorithm. The LRTs for the three sets of hypotheses in Eq. (6.7) are computed using the binary log-likelihood function $\ell(\beta)$ given in Section 3.2 based on the maximum likelihood estimators, β° under H_0 , $\tilde{\beta}$ under H_1 , and $\hat{\beta}$ under H_2 . Consider a testing problem for H_0 against $H_2 - H_0$. Then the unrestricted LRT statistic is given by

$$T_{02} = 2[\ell(\hat{\beta}) - \ell(\beta^\circ)],$$

which asymptotically follows $\chi^2(r)$ under the null hypothesis H_0 . If T_{02} is large, then the unrestricted test rejects H_0 in favor of $H_2 - H_0$. When testing H_1 against $H_1 - H_0$, (i.e. when the parameter space is restricted by H_1), the restricted LRT statistic is given by

$$T_{01} = 2[\ell(\tilde{\beta}) - \ell(\beta^\circ)].$$

When testing H_1 against $H_2 - H_1$, the restricted LRT statistic is given by

$$T_{12} = 2[\ell(\hat{\beta}) - \ell(\tilde{\beta})].$$

When H_1 is true, the usefulness of the test related to the restricted LRT statistic T_{01} can be confirmed by performing the goodness of fit test, which rejects H_1 for large values of the restricted LRT statistic T_{12} . In generalized linear models, the asymptotic distributions for constrained likelihood ratio tests T_{01} and T_{12} are demonstrated to be chi-bar-square, as outlined in the following theorem.

Theorem 6.1 (Asymptotic LRT distribution when $\mathbf{Y}$ belongs to FE): *Let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^p$ and $\mathbf{V}(\theta_0)$ be a $p \times p$ positive definite matrix. Then under the null hypothesis*

H_0 , the asymptotic distributions of the LRT statistics $T_{01}(\mathbf{V}(\boldsymbol{\theta}_0), \mathcal{C})$ and $T_{12}(\mathbf{V}(\boldsymbol{\theta}_0), \mathcal{C})$ are given as follows:

$$\lim_{n \rightarrow \infty} P_{\boldsymbol{\theta}_0}\{T_{01} \geq c\} = \sum_{i=0}^q w_i(q, \mathbf{A}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{A}^T, \mathcal{C})P(\chi_i^2 \geq c), \quad (6.8)$$

$$\lim_{n \rightarrow \infty} P_{\boldsymbol{\theta}_0}\{T_{12} \geq c\} = \sum_{i=0}^q w_{q-i}(q, \mathbf{A}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{A}^T, \mathcal{C})P(\chi_i^2 \geq c), \quad (6.9)$$

for any $c \geq 0$, where q is the rank of $\mathbf{A}$, $\boldsymbol{\theta}_0$ is a true value of $\boldsymbol{\theta}$ under H_0 , $\mathbf{V}(\boldsymbol{\theta}_0)$ is the inverse of the fisher information matrix, $w_i(q, \mathbf{A}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{A}^T)$ are some non-negative numbers and $\sum_{i=0}^q w_i(q, \mathbf{A}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{A}^T) = 1$.

Proof. Given the binary log-likelihood function in Section 3.2 and since the likelihood is the product of functions in the exponential family which are known to be concave and bounded, the assumptions of uniqueness, differentiable (i.e. the first three partial derivatives of $\log f(\mathbf{y}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ exist almost everywhere) and finite variance (i.e. the Fisher information matrix, $I(\boldsymbol{\theta})$, is finite and positive definite) are satisfied. From simulation results, it was noted that the unconstrained values of $\hat{\boldsymbol{\theta}}$ were normally distributed under the models considered. Let $\boldsymbol{\theta}_0$ denote its true model parameter value. For large samples, the following four axioms properties about $\hat{\boldsymbol{\theta}}$ hold when there are no inequality constraints on $\boldsymbol{\theta}$:

- (1) $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$ almost surely as $n \rightarrow \infty$,
- (2) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightarrow N(\mathbf{0}, \mathbf{V})$, in distribution,
- (3) $\hat{\mathbf{V}}$ is a consistent estimator of $\mathbf{V}$.

The LRT for testing $H_0 : \mathbf{A}\boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \mathbf{A}\boldsymbol{\theta} \neq \mathbf{0}$ is asymptotically χ_r^2 under H_0 where $r = \text{rank}(\mathbf{A})$.

Let $q(\boldsymbol{\theta}) = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{V}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\mathbf{V}}^2$ and $\hat{q}(\boldsymbol{\theta}) = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\hat{\mathbf{V}}}^2$.

Define

$$\hat{T} = \min_{\boldsymbol{\theta} \in \mathcal{M}} \{\hat{q}(\boldsymbol{\theta})\} - \min_{\boldsymbol{\theta} \in \mathcal{C}} \{\hat{q}(\boldsymbol{\theta})\}, \quad (6.10)$$

If we replace $\hat{\mathbf{V}}^{-1}$ by its probability limit, $\mathbf{V}$, then it follows that $|q(\boldsymbol{\theta}) - \hat{q}(\boldsymbol{\theta})| = o_p(1)$. Let $\mathbf{X}_n = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $\mathbf{X} \sim N(\mathbf{0}, \mathbf{V})$. From Theorem 5.3 suppose that $\boldsymbol{\theta} \in \mathcal{M}$ and from Lemma 6.1 suppose that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. Then we have

$$\begin{aligned} \hat{T} &= \min_{\boldsymbol{\theta} \in \mathcal{M}} \{(\mathbf{X}_n - \boldsymbol{\theta})^T \mathbf{V}(\mathbf{X}_n - \boldsymbol{\theta})\} - \min_{\boldsymbol{\theta} \in \mathcal{C}} \{(\mathbf{X}_n - \boldsymbol{\theta})^T \mathbf{V}(\mathbf{X}_n - \boldsymbol{\theta})\} + o_p(1) \\ &\xrightarrow{d} \min_{\boldsymbol{\theta} \in \mathcal{M}} \{(\mathbf{X} - \boldsymbol{\theta})^T \mathbf{V}(\mathbf{X} - \boldsymbol{\theta})\} - \min_{\boldsymbol{\theta} \in \mathcal{C}} \{(\mathbf{X} - \boldsymbol{\theta})^T \mathbf{V}(\mathbf{X} - \boldsymbol{\theta})\}. \end{aligned} \quad (6.11)$$

Since $q(\boldsymbol{\theta}) = -2\ell(\boldsymbol{\theta})$, the likelihood ratio test statistics are written as

$$T_{01} = q(\boldsymbol{\theta}^0) - q(\tilde{\boldsymbol{\theta}}) \text{ and } T_{12} = q(\tilde{\boldsymbol{\theta}}) - q(\hat{\boldsymbol{\theta}}), \quad (6.12)$$

where $\boldsymbol{\theta}^0$ and $\tilde{\boldsymbol{\theta}}$ are estimators of $\boldsymbol{\theta}$ under H_0 and H_1 , respectively. Recall the parameter space under H_0 , H_1 and H_2 as $\mathcal{C}_0 = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{A}\boldsymbol{\theta} = \mathbf{b}\}$, $\mathcal{C}_1 = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{A}\boldsymbol{\theta} \leq \mathbf{b}\}$, and $\mathcal{C}_2 = \{\boldsymbol{\theta} \in \mathbb{R}^p : \text{no restriction on } \boldsymbol{\theta}\}$, respectively. Similar to the normal case, the least squares projections $P_{\hat{\mathbf{V}}}(\hat{\boldsymbol{\theta}}|\mathcal{C}_0)$ and $P_{\hat{\mathbf{V}}}(\hat{\boldsymbol{\theta}}|\mathcal{C}_1)$ are the restricted estimators obtained by minimizing $\hat{q}(\boldsymbol{\theta}) = \|\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|_{\hat{\mathbf{V}}}^2$ under H_0 and H_1 , respectively. Thus we may approximate the LRT T_{01} in (6.12) with $o_p(1)$ (see Appendix D) error as

$$\begin{aligned} T_{01} &= n\|\hat{\boldsymbol{\theta}} - P_{\hat{\mathbf{V}}}(\hat{\boldsymbol{\theta}}|\mathcal{C}_0)\|_{\hat{\mathbf{V}}}^2 - n\|\hat{\boldsymbol{\theta}} - P_{\hat{\mathbf{V}}}(\hat{\boldsymbol{\theta}}|\mathcal{C}_1)\|_{\hat{\mathbf{V}}}^2 + o_p(1) \\ &= n\|P_{\hat{\mathbf{V}}}(\hat{\boldsymbol{\theta}}|\mathcal{C}_1) - P_{\hat{\mathbf{V}}}(\hat{\boldsymbol{\theta}}|\mathcal{C}_0)\|_{\hat{\mathbf{V}}}^2 + o_p(1). \end{aligned}$$

Due to properties of the least squares projection, for any $\boldsymbol{\theta}_0 \in \mathcal{C}_0$ we have that

$$\sqrt{n}(P_{\hat{\mathbf{V}}}(\hat{\boldsymbol{\theta}}|\mathcal{C}_i) - \boldsymbol{\theta}_0) = P_{\hat{\mathbf{V}}}(\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)|\mathcal{C}_i) \quad \forall i = 0, 1.$$

Thus, if $\boldsymbol{\theta}_0 \in \mathcal{C}_0$ is the underlying model parameter under H_0 , then by the above three axioms and the continuity property of the projection operator,

$$\begin{aligned}
T_{01} &= \|P_{\hat{\mathbf{V}}}(\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)|\mathcal{C}_1) - P_{\hat{\mathbf{V}}}(\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)|\mathcal{C}_0)\|_{\hat{\mathbf{V}}}^2 + o_p(1) \\
&\xrightarrow{d} \|P_{\hat{\mathbf{V}}}(\mathbf{Z}|\mathcal{C}_1) - P_{\hat{\mathbf{V}}}(\mathbf{Z}|\mathcal{C}_0)\|_{\hat{\mathbf{V}}}^2,
\end{aligned}$$

where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{V})$ and from Theorem 5.3, the asymptotic null distribution of LRT T_{01} is chi-bar-square with weights $w_i(r, \mathbf{A}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{A}^T)$. Similarly the asymptotic null distribution of LRT T_{12} under H_0 may be derived. Applying a large sample approximation to T_{12} in Eq. (6.12), we have

$$T_{12} \xrightarrow{d} \|\mathbf{Z} - P_{\hat{\mathbf{V}}}(\mathbf{Z}|\mathcal{C}_1)\|_{\hat{\mathbf{V}}}^2,$$

for $\boldsymbol{\theta}_0 \in \mathcal{C}_0$. The asymptotic distribution of LRT T_{12} with a parameter $\boldsymbol{\theta}_0$ in H_0 may be immediately obtained by Theorem 5.3. $\square$

6.3.1 Empirical Results for Restricted LRT Under Binary GLM

To investigate the performance of the proposed GP algorithm for GLM, a simulation study was conducted using two different sample sizes $n = 100$ and $n = 300$. The powers of restricted (T_{01}, T_{12}) and unrestricted T_{02} LRT described in Theorem 6.1 were computed.

We use R for performing the constrained statistical inference under the GLM. The covariate matrix and the two constraints imposed on the parameters are the same as those used in Section 6.2.1.

$$\begin{aligned}
\beta_0 + \beta_1 + \beta_2 &\leq 0 \\
-\beta_0 + \beta_1 - \beta_2 &\leq 5
\end{aligned}
\Rightarrow \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 5 \end{pmatrix} \quad (6.13)$$

In the instance of restricted LRT under binary GLM, we estimate $\boldsymbol{\beta}$ under two cases:

- (a) H_0 , where both constraints are active $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ (i.e. both hold with equality). This case is used to test for empirical sizes (i.e. the significance level).

(b) H_1 , where at least one constraint is inactive $\mathbf{a}_i^T \boldsymbol{\beta} < b_i$ (i.e. at least one is a strict inequality). This case is used to test for the empirical power of the test.

For case (a), we use three sets of parameters whereas in case (b), we use four sets to satisfy H_1 , where the first two sets use the first constraint as an active constraint, and the second two sets use the second constraint as an active constraint. For more details see Table 6.5.

Table 6.5: The empirical powers and sizes of restricted and unrestricted LRT for $n = 100$ and $n = 300$ at 5% significance level

LRT power for Bernoulli Model								
Sample Sizes		n = 100				n = 300		
Case	$\boldsymbol{\beta}^T$	T_{01}	T_{02}	T_{12}		T_{01}	T_{02}	T_{12}
a	(-2.00,2.50,-0.50)	3.6	5.4	8.1		4.4	4.8	6.0
	(-2.20,2.50,-0.30)	4.4	7.3	8.6		4.1	5.3	5.9
	(-2.25,2.50,-0.25)	4.1	7.1	8.7		4.7	5.4	5.3
b	(-1.75,2.25,-0.50)	12.6	8.3	4.7		20.1	12.0	2.8
	(-1.75,2.00,-0.25)	29.1	17.2	2.5		57.8	41.0	1.5
	(-2.25,2.30,-0.45)	44.3	28.1	2.1		83.4	66.4	3.1
	(-2.25,2.20,-0.55)	73.1	59.4	2.2		99.4	98.1	2.0

Table 6.5 shows empirical levels and powers of LRTs at 5% level of significance based on 1000 replications. We obtained the restricted estimates using the GP algorithm described in Section 6.2. For testing the restricted hypothesis H_0 against $H_1 - H_0$, we use the LRT statistic T_{01} . For testing the restricted goodness-of-fit H_1 against $H_2 - H_1$, we use the LRT statistic T_{12} , and for testing H_0 against $H_2 - H_0$, we use the LRT statistic T_{02} .

As stated above in item (a), the empirical sizes are close to the nominal 5% significance level; however, when the sample size is increased from 100 to 300, the empirical size becomes even closer to the nominal significance level, and by extension, we notice less bias in empirical sizes. Additionally, as stated in item (b), T_{01} and T_{02} provide the empirical powers for the restricted

and unrestricted tests. When testing the restricted LRT H_1 against $H_2 - H_1$, the values of β are within H_1 ; thus the values of T_{12} in case (b) represent the size, and not the power. Also, these values are smaller than the 5% significance level since the particular null value is not the least favourable one. The restricted test T_{01} demonstrates acceptable empirical size and significantly better power performance than T_{02} when the constraints are satisfied. These findings (using LRT T_{01} when constraints exist) are consistent with the comparison of population order mean as seen in Section 5.2, where we use the restricted F -test rather than the standard F -test.

6.4 GP Algorithm for Multinomial Logit Model

Let the convex cone $\mathcal{C}_1 = \{\beta : \mathbf{A}\beta \leq \mathbf{b}\}$ denote the constrained parameter space, where $\mathbf{A}$ is a $r \times p$ matrix of full row rank, ($r < p$), p is the dimension of $\beta = (\beta_1^T, \beta_2^T, \dots, \beta_{c-1}^T)^T$, and β_j are the estimates for the j^{th} category. In order to maximize the multinomial log-likelihood function under such inequality constraints, we implement a modified version of the gradient projection algorithm of Jamshidian (2004) (for more information see Theorem 4.2), which searches active constraint sets to determine the optimal solution. Another modification to the GP algorithm that was implemented is in Step 4 of the aforementioned theorem; it computes the scalar α from $\max_{\alpha} \{\ell(\beta + \alpha \mathbf{d}) : 0 \leq \alpha \leq \infty\}$. This was done by using a line search step-halving to obtain the smallest integer $k \geq 0$ such that $\ell(\beta_{new}) > \ell(\beta_{old})$ for $\beta_{new} = \beta_{old} + 0.5^k \mathbf{d}$, where $\mathbf{d}$ is the feasible direction vector (4.19). Therefore, when k is obtained, 0.5^k is used as an initial starting point in the built-in R function, `optim`, to find the optimal α value that maximizes the multinomial loglikelihood $\ell(\beta + \alpha \mathbf{d})$. To examine the performance of the GP algorithm for the multinomial logit model, we conducted simulations using fictitious information.

6.4.1 Simulation Study

To perform simulations with the multinomial logit model under linear constraints, we use a fictitious Canadian political election study, in which voter intent is studied based on two random

covariates for the Conservatives (CP), Liberals (LP), and Other parties. The parameter vectors $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \beta_{21})^T$ and $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12}, \beta_{22})^T$ represent the coefficient for CP and LP categories, respectively. The third category (Other parties) was used as a base reference category. Similar to the binary GLM, the explanatory variables $\mathbf{x}_i^T = (x_{1i}, x_{2i})$ were generated from a bivariate normal distribution with mean $\boldsymbol{\mu} = (2, 1)^T$ and variance-covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}$. For the purpose of this thesis, we identify four constraints, two per category as follows:

$$\begin{aligned} \beta_{01} + \beta_{11} + \beta_{21} &\leq 0.0 \\ -\beta_{01} + \beta_{11} - \beta_{21} &\leq 5.0 \\ -\beta_{02} + \beta_{12} + \beta_{22} &\leq 0.5 \\ \beta_{02} - \beta_{12} + 2\beta_{22} &\leq 4.0 \end{aligned} \Rightarrow \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 & 2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0.0 \\ 5.0 \\ 0.5 \\ 4.0 \end{pmatrix} \quad (6.14)$$

Matrix $\mathbf{A}$ is a block diagonal matrix with the blocks on the diagonal representing the constraints for each category. If we have both equality and inequality constraints, the matrix $\mathbf{A}$ is represented as $\begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$, where $\mathbf{A}_1$ and $\mathbf{A}_2$ are block matrices representing equality and inequality constraints, respectively. The study was conducted by generating 1000 datasets from multinomial distributions using three different sample sizes (350, 700, and 1000). Section 6.5 covers the restricted MLE for multinomial logit and Section 6.6 covers the restricted test simulation results.

6.5 Restricted MLE for Multinomial Logit Using GP

To compute the restricted MLEs for the multinomial generalized logit, we use the same GP approach as the GLM binary restricted ML method in Section 6.2, with one minor difference in which the non-restricted MLEs are computed using the NR algorithm, implemented in R by myself. Using both restricted and unrestricted methods, we compare the simulated sampled mean, bias and mean square error (MSE) of the estimates of parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$.

We examine several types of preliminary points $\boldsymbol{\beta}$, where:

- (a) both constraints/prior-knowledge are active $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ (i.e. both hold with equality), for each of the two categories,
- (b) at least one constraint is inactive $\mathbf{a}_i^T \boldsymbol{\beta} < b_i$ (i.e. at least one is a strict inequality), in one of the two categories:
 - $\mathbf{b}_1$: The first constraint in equation (6.13) is inactive for each of the two categories.
 - $\mathbf{b}_2$: The second constraint in equation (6.14) is inactive for each of the two categories.
- (c) both constraints are inactive $\mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}$ (i.e. both are within the constraint cone) for each of the two categories.

Based on our existing knowledge of normal models [15] and [10] and binary GLM models (Section 6.2.1) with order-constrained inference, we would expect the restricted estimators to have a larger bias and a smaller MSE, particularly when the values of $\boldsymbol{\beta}$ are close to the boundary. In these instances, the unrestricted estimators are likely to be found outside the cone. Since the estimates are always forced toward the cone by projecting the estimators onto the closed convex cone, we end up with a bias. The results for unrestricted and restricted multinomial logit models are shown in Tables 6.6 to 6.8 for samples of sizes 350, 700 and 1000.

Table 6.6: Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for MN Logit Model ($N = 350$)

Case	Parameter	True Value	Unrestricted			Restricted		
			Mean	Bias	MSE	Mean	Bias	MSE
a	β_{01}	-2.00	-2.0003	-0.0003	0.3231	-1.8678	0.1322	0.0534
	β_{11}	2.50	2.5411	0.0411	0.1905	2.3808	-0.1192	0.0330
	β_{21}	-0.50	-0.3483	0.1517	0.6980	-0.5924	-0.0924	0.0394
	β_{02}	0.50	0.5520	0.0520	0.2210	0.4375	-0.0625	0.0662
	β_{12}	-0.50	-0.5263	-0.0263	0.1812	-0.4800	0.0200	0.0556
	β_{22}	1.50	1.6184	0.1184	0.9566	1.1009	-0.3991	0.3539
b ₁	β_{01}	-2.50	-2.5623	-0.0623	0.8351	-2.0975	0.4025	0.5530
	β_{11}	1.25	1.2577	0.0077	0.3917	0.9656	-0.2845	0.2709
	β_{21}	-1.25	-1.3420	-0.0920	0.7691	-0.9502	0.2998	0.5114
	β_{02}	1.70	1.7492	0.0492	0.1729	1.5632	-0.1368	0.0958
	β_{12}	-0.30	-0.3225	-0.0225	0.1028	-0.1740	0.1260	0.0563
	β_{22}	1.00	1.0781	0.0781	0.1997	0.9208	-0.0792	0.0556
b ₂	β_{01}	-2.00	-2.0519	-0.0519	0.3433	-2.2186	-0.2186	0.2119
	β_{11}	1.00	1.0298	0.0298	0.1827	1.1456	0.1456	0.1084
	β_{21}	1.00	1.0664	0.0664	0.3399	0.8255	-0.1745	0.1307
	β_{02}	-1.00	-1.0297	-0.0297	0.1477	-1.0113	-0.0113	0.1157
	β_{12}	0.50	0.5134	0.0134	0.0816	0.4872	-0.0128	0.0618
	β_{22}	-1.00	-1.0726	-0.0726	0.3959	-1.2275	-0.2275	0.3945
c	β_{01}	-2.00	-2.0713	-0.0713	0.3241	-2.0603	-0.0603	0.2875
	β_{11}	0.50	0.5404	0.0404	0.1716	0.5331	0.0331	0.1531
	β_{21}	0.50	0.4751	-0.0250	0.2117	0.4790	-0.0210	0.1925
	β_{02}	-0.50	-0.5163	-0.0163	0.1476	-0.4773	0.0227	0.1120
	β_{12}	-0.50	-0.5036	-0.0036	0.0994	-0.5317	-0.0317	0.0808
	β_{22}	0.20	0.1472	-0.0528	0.2251	0.1603	-0.0397	0.2179

Table 6.7: Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for MN Logit Model ($N = 700$)

Case	Parameter	True Value	Unrestricted			Restricted		
			Mean	Bias	MSE	Mean	Bias	MSE
a	β_{01}	-2.00	-1.9842	0.0158	0.1312	-1.9018	0.0982	0.0285
	β_{11}	2.50	2.5037	0.0037	0.0770	2.4069	-0.0931	0.0193
	β_{21}	-0.50	-0.4217	0.0784	0.2678	-0.5682	-0.0682	0.0207
	β_{02}	0.50	0.5057	0.0057	0.0919	0.4410	-0.0590	0.0323
	β_{12}	-0.50	-0.5041	-0.0041	0.0781	-0.4789	0.0211	0.0273
	β_{22}	1.50	1.5375	0.0375	0.3675	1.2233	-0.2767	0.1660
b ₁	β_{01}	-2.50	-2.5313	-0.0313	0.3987	-2.1718	0.3282	0.4494
	β_{11}	1.25	1.2614	0.0114	0.1841	1.0310	-0.2190	0.2044
	β_{21}	-1.25	-1.2848	-0.0348	0.2586	-0.9797	0.2703	0.3691
	β_{02}	1.70	1.7290	0.0290	0.0846	1.5730	-0.1270	0.0856
	β_{12}	-0.30	-0.3160	-0.0160	0.0509	-0.1908	0.1092	0.0473
	β_{22}	1.00	1.0315	0.0315	0.0946	0.9240	-0.0760	0.0333
b ₂	β_{01}	-2.00	-2.0249	-0.0249	0.1509	-2.1415	-0.1415	0.0971
	β_{11}	1.00	1.0182	0.0182	0.0804	1.0988	0.0988	0.0519
	β_{21}	1.00	1.0289	0.0289	0.1419	0.8682	-0.1318	0.0662
	β_{02}	-1.00	-1.0106	-0.0106	0.0713	-0.9948	0.0052	0.0571
	β_{12}	0.50	0.5064	0.0064	0.0400	0.4855	-0.0145	0.0298
	β_{22}	-1.00	-1.0116	-0.0116	0.1763	-1.1239	-0.1239	0.1736
c	β_{01}	-2.00	-2.0054	-0.0054	0.1516	-2.0005	-0.0005	0.1499
	β_{11}	0.50	0.4958	-0.0042	0.0835	0.4923	-0.0077	0.0828
	β_{21}	0.50	0.5181	0.0181	0.1026	0.5204	0.0204	0.1021
	β_{02}	-0.50	-0.4961	0.0039	0.0675	-0.4787	0.0213	0.0567
	β_{12}	-0.50	-0.5088	-0.0088	0.0453	-0.5215	-0.0215	0.0399
	β_{22}	0.20	0.2090	0.0090	0.0947	0.2157	0.0157	0.0937

Table 6.8: Simulated Mean, Bias and MSE of Restricted and Unrestricted Estimates for MN Logit Model ($N = 1000$)

Case	Parameter	True Value	Unrestricted			Restricted		
			Mean	Bias	MSE	Mean	Bias	MSE
a	β_{01}	-2.00	-2.0227	-0.0227	0.1120	-1.9196	0.0804	0.0201
	β_{11}	2.50	2.5291	0.0291	0.0660	2.4236	-0.0765	0.0137
	β_{21}	-0.50	-0.4890	0.0110	0.1964	-0.5641	-0.0641	0.0169
	β_{02}	0.50	0.5039	0.0039	0.0753	0.4589	-0.0411	0.0262
	β_{12}	-0.50	-0.4999	0.0001	0.0606	-0.4843	0.0157	0.0210
	β_{22}	1.50	1.4898	-0.0102	0.2775	1.2547	-0.2453	0.1257
b ₁	β_{01}	-2.50	-2.5238	-0.0238	0.2510	-2.2287	0.2713	0.3720
	β_{11}	1.25	1.2610	0.0110	0.1178	1.0722	-0.1778	0.1742
	β_{21}	-1.25	-1.2694	-0.0194	0.1751	-1.0063	0.2438	0.3295
	β_{02}	1.70	1.7233	0.0233	0.0562	1.5781	-0.1219	0.0817
	β_{12}	-0.30	-0.3115	-0.0115	0.0341	-0.1958	0.1042	0.0465
	β_{22}	1.00	1.0304	0.0304	0.0607	0.9403	-0.0597	0.0233
b ₂	β_{01}	-2.00	-2.0044	-0.0044	0.1064	-2.1054	-0.1054	0.0646
	β_{11}	1.00	1.0028	0.0028	0.0569	1.0727	0.0727	0.0344
	β_{21}	1.00	1.0315	0.0315	0.0971	0.8955	-0.1045	0.0445
	β_{02}	-1.00	-1.0142	-0.0142	0.0480	-1.0006	-0.0006	0.0389
	β_{12}	0.50	0.5055	0.0055	0.0274	0.4876	-0.0124	0.0210
	β_{22}	-1.00	-1.0256	-0.0256	0.1315	-1.1219	-0.1219	0.1274
c	β_{01}	-2.00	-2.0195	-0.0195	0.1089	-2.0168	-0.0168	0.1082
	β_{11}	0.50	0.5103	0.0103	0.0610	0.5084	0.0084	0.0607
	β_{21}	0.50	0.4953	-0.0048	0.0766	0.4965	-0.0035	0.0764
	β_{02}	-0.50	-0.5080	-0.0080	0.0498	-0.4983	0.0017	0.0440
	β_{12}	-0.50	-0.4995	0.0005	0.0322	-0.5065	-0.0065	0.0293
	β_{22}	0.20	0.1793	-0.0207	0.0691	0.1830	-0.0171	0.0686

As expected, we notice from Tables 6.6 to 6.8 that in almost all cases the restricted MLE have a larger bias and smaller MSE than their unrestricted counterparts. For case (c), the pattern still holds true; however, we notice that the results are very close in values as one could expect given that the true values are from within the cone. Additionally, with an increasing sample size from 350 to 700 to 1000, the asymptotic bias and MSE significantly decrease for both restricted and unrestricted MLEs. We note that for restricted estimates when the true parameter β is chosen closer to the boundary points of the convex cone, the bias increases and the MSE decreases.

The usual unrestricted MLE is not necessarily suitable when the generating/preliminary point β is not within the cone as it does not incorporate prior knowledge given by (6.13). This means that we cannot determine a targeted direction for the parameters. Table 6.9 shows the percentage of the unrestricted estimates that lie within the cone (i.e. satisfy the constraints). According to Remark 6.1, the restricted MLE is considered the same as the unrestricted MLE in such cases.

Table 6.9: Percentage of unrestricted MLE that satisfy the constraints

% of Restricted MLEs = Unrestricted MLEs				
Cases	a	b ₁	b ₂	c
$N = 350$	3.7%	16.4%	23.3%	71.5%
$N = 700$	3.4%	17.2%	24.6%	82.9%
$N = 1000$	3.4%	15.5%	24.4%	89.0%

As mentioned earlier in this thesis, the properties of the unrestricted MLEs do not hold true for cases (b₁) and (b₂), where the distributions for the 1000 replications for restricted MLEs are not normal. Please refer to the distributions shown in Appendix G, Sections G.1 and G.2. However, for cases (a) and (c), the distributions of the 1000 replications of the restricted MLEs appear normal for all sample sizes $N = 350, 700$ and 1000.

6.6 Restricted Tests for Multinomial Logit

We also study the empirical properties of restricted LRTs using the same simulation setting as described in Section 6.5. The powers of restricted (T_{01}, T_{12}) and unrestricted T_{02} LRTs described in Theorem 6.1 were computed. We use the statistical software R, where a few functions were written to analyse the data. The covariate matrix and the four constraints (two per category) imposed on the multinomial logit parameters are the same as per Section 6.4.1:

$$\begin{aligned} \beta_{01} + \beta_{11} + \beta_{21} &\leq 0.0 \\ -\beta_{01} + \beta_{11} - \beta_{21} &\leq 5.0 \\ -\beta_{02} + \beta_{12} + \beta_{22} &\leq 0.5 \\ \beta_{02} - \beta_{12} + 2\beta_{22} &\leq 4.0 \end{aligned} \Rightarrow \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 & 2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0.0 \\ 5.0 \\ 0.5 \\ 4.0 \end{pmatrix} \quad (6.15)$$

In the instance of restricted LRT under multinomial logit, we estimate $\boldsymbol{\beta}$ under two cases:

- (a) H_0 , where both constraints are active $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ (i.e. both hold with equality) for each of the two categories. This case is used to test for empirical sizes (i.e. the significance level).
- (b) H_1 , where at least one constraint is inactive $\mathbf{a}_i^T \boldsymbol{\beta} < b_i$ (i.e. at least one is a strict inequality) for each of the two categories. This case is used to find the empirical power of the test.

For case (a), we use three sets of parameters whereas in case (b), we use four sets to satisfy H_1 where the first two sets use the first constraint as an active constraint and the second two sets use the second constraint as an active constraint.

Table 6.10 below shows empirical levels and powers of LRTs for $N = 250$, $N = 350$, $N = 700$ and $N = 1000$ at 5% level of significance for 1000 replications. We obtained the restricted estimates using the GP algorithm described above. For testing the restricted hypothesis H_0 against $H_1 - H_0$, we use the LRT statistic T_{01} . For testing the restricted goodness-of-fit H_1 against $H_2 - H_1$, we use the LRT statistic T_{12} , and for testing the unrestricted likelihood ratio test H_0 against $H_2 - H_0$, we use the LRT statistic T_{02} .

Table 6.10: Empirical powers and sizes of restricted and unrestricted LRT for $N = (250, 350, 700, \text{ and } 1000)$ at 5% significance level

LRT for Multinomial Logit Model														
Case	Parameters of Categories		Sample Sizes											
			N = 250			N = 350			N = 700			N = 1000		
	β_1^T	β_2^T	T_{01}	T_{02}	T_{12}	T_{01}	T_{02}	T_{12}	T_{01}	T_{02}	T_{12}	T_{01}	T_{02}	T_{12}
a	(-2.00, 2.50, -0.50)	(0.50, -0.50, 1.50)	4.0	4.9	6.3	4.8	5.7	6.8	4.6	4.7	4.9	4.5	4.5	4.9
	(-2.20, 2.50, -0.30)	(1.50, 0.50, 1.50)	4.9	5.2	6.5	5.1	5.0	6.7	5.6	4.8	5.5	4.1	3.5	5.1
	(-2.25, 2.50, -0.25)	(1.24, 0.24, 1.50)	4.2	4.4	5.6	4.8	4.8	5.7	4.9	5.1	5.7	5.2	3.9	6.0
b	(-1.75, 2.25, -0.50)	(-1.00, 0.50, -1.00)	84.4	75.4	2.0	93.6	87.4	1.6	99.8	99.6	2.6	100.0	100.0	1.4
	(-1.75, 2.00, -0.25)	(1.00, 0.50, 1.00)	38.0	26.3	2.5	46.3	34.4	1.6	79.2	67.8	1.1	91.7	84.3	2.2
	(-2.25, 2.30, -0.45)	(1.70, -0.30, 1.00)	59.3	50.6	5.0	74.5	64.0	3.5	96.9	93.3	2.9	99.6	98.9	3.8
	(-2.25, 2.20, -0.55)	(1.20, -0.80, 1.00)	60.6	48.2	4.1	72.4	62.0	3.7	96.2	92.1	4.8	98.8	98.6	3.9

As stated above in item (a), the empirical sizes are close to nominal 5% significance level; however, when the sample size is increased from 250 to 1000, the empirical size becomes even closer to the significance level, and by extension, we notice less bias in empirical sizes. Additionally, as stated in item (b), T_{01} and T_{02} provide the empirical powers for the restricted and unrestricted tests. When testing the restricted LRT H_1 against $H_2 - H_1$, the values of β are within H_1 ; thus the values of T_{12} in case (b) represent the size, not the power. Also, these values are smaller than the 5% significance level since the particular null value is not the least favourable one. The restricted test T_{01} demonstrates acceptable empirical size and significantly better power performance than T_{02} when the constraints are satisfied. These findings (using LRT T_{01} when constraints exist) are consistent with the comparison of population ordered mean as seen in Section 5.2 where we use the restricted F -test rather than the standard F -test and the restricted LRT for GLM in Section 6.3.1.

Chapter 7

Applications - Analysing CCHS Data Using Restricted Multinomial Logit

7.1 Canadian Community Health Survey

The Canadian Community Health Survey (CCHS) was originally developed in 1997 as a collaborative work between the Canadian Institute for Health Information (CIHI), Health Canada and Statistics Canada to address gaps and issues with the health information system. The CCHS aims to collect health-related information at the health region level, whereby health regions are defined by the provinces and represent administrative areas or regions of interest to health authorities [29]. The CCHS aims to 1) support health surveillance programs; 2) provide a single data source for health research on small populations and rare characteristics; 3) release easily accessible information to a diverse community of users in a timely manner; and 4) create a flexible survey instrument that includes a rapid response option to address emerging issues related to the health of the population [30]. Originally launched in 2000, and every two-years thereafter, the CCHS modified its reporting in 2007 and began collecting data and publishing its results every year. An extensive redesign of the survey methodology was completed in 2015, and so the data being used for the purposes of this thesis will be 2012 data, which will give us the flexibility of building on the existing works of Karelyn Davis [10] and Predrag Mizdrak [101]. The survey covers Canada's provinces and territories, surveying persons living in private dwellings aged 12 years and older. It does not include, however, persons living on Indian Reserves or Crown lands, persons residing in institutions, full-time members of the Canadian Forces, and residents of some remote regions. Despite the persons excluded from the survey,

the CCHS still covers approximately 98% of the Canadian population aged 12 and older [31].

7.2 Description of the Asthma Subset of CCHS Data

For the purposes of this thesis and to demonstrate the unrestricted and restricted methods for the multinomial logit model, a subset of the CCHS data is used, which only includes respondents who indicated they were asthmatic. We use the response variable (CHPGMDC), which represents the number of consultations the respondents had over a year with medical doctors. We define three levels, which determine if the respondent is a light user of the health system (0 to 5 consultations), a moderate user of the health system (5 to 10 consultations), or a heavy user of the health system (10 or more consultations). Our goal is to assess the relationships between our response variable and its predictors (Age, Sex, Smoker, Symptoms) and how these may impact the number of consultations with a medical doctor. Our predictors are organised as follows:

- Age (DHHGAGE) is coded into four binary covariates:
 - Age1, representing ages 12 to 24 years old coded as (1), otherwise (0);
 - Age2, representing ages 25 to 44 years old coded as (1), otherwise (0);
 - Age3, representing ages 45 to 64 years old coded as (1), otherwise (0); and
 - Age4, representing ages 65 years old and older coded as (1), otherwise (0).
- Sex (DHHSEX): coded as (0) for Male, (1) for Female.
- Smoker (SMK_202): coded as (0) for Non-Smoker (if respondent does not smoke at all), (1) for Smoker (if respondent smokes daily or occasionally).
- Symptoms (CCC_035): is coded as (0) for No Symptoms (if the respondent indicated not having asthma symptoms or attack in the past 12 months), (1) for Symptoms (if the respondent indicated experiencing asthma symptoms or an attack in the past 12 months).

The asthma subset of data from CCHS contains 5305 records of raw data. For our analysis, we removed 135 records where the responses were categorized as “Not Applicable”, “Refusal”, “Not Stated”, or “Don’t know”. These can be broken down by each of the variables as follows: 86 records were removed from the CHPGMDC; 18 records were removed from CCC_035; and 31 records were removed from SMK_202. As a result, our subset contains 5170 records. Table 7.1 summarizes the data described above.

Table 7.1: Asthma Data Subset from Canadian Community Health Survey 2012

Covariates						Count			
age1	age2	age3	sex	smoker	symptom	Light	Moderate	Heavy	Total
1	0	0	0	0	0	251	29	11	291
1	0	0	0	0	1	100	32	13	145
1	0	0	0	1	0	66	7	5	78
1	0	0	0	1	1	25	5	0	30
1	0	0	1	0	0	177	36	22	235
1	0	0	1	0	1	124	41	33	198
1	0	0	1	1	0	35	16	9	60
1	0	0	1	1	1	28	15	11	54
0	1	0	0	0	0	130	16	6	152
0	1	0	0	0	1	113	28	11	152
0	1	0	0	1	0	53	7	12	72
0	1	0	0	1	1	51	14	5	70
0	1	0	1	0	0	134	54	41	229
0	1	0	1	0	1	166	82	89	337
0	1	0	1	1	0	65	24	24	113
0	1	0	1	1	1	58	28	44	130
0	0	1	0	0	0	140	30	14	184
0	0	1	0	0	1	104	54	23	181
0	0	1	0	1	0	46	15	4	65
0	0	1	0	1	1	36	13	15	64
0	0	1	1	0	0	199	78	50	327
0	0	1	1	0	1	243	121	100	464
0	0	1	1	1	0	57	24	18	99
0	0	1	1	1	1	65	40	54	159
0	0	0	0	0	0	135	38	37	210
0	0	0	0	0	1	76	59	41	176
0	0	0	0	1	0	18	5	7	30
0	0	0	0	1	1	14	7	6	27
0	0	0	1	0	0	230	89	52	371
0	0	0	1	0	1	162	115	82	359
0	0	0	1	1	0	32	17	12	61
0	0	0	1	1	1	29	9	9	47

Table 7.2 contains the summary statistics for the asthma data from the CCHS. It shows the important covariates (age, sex, smoker, symptoms) and their effect on health system use (light, moderate, heavy) using unweighted counts.

Table 7.2: Summary Statistics for Canadian Community Health Survey 2012

CCHS Data		Response: # of consultations						
		Count				Percentage		
Covariate	Level	Light	Moderate	Heavy	Total	Light	Moderate	Heavy
age	1 = 12-24	806	181	104	1091	73.9%	16.6%	9.5%
	2 = 25-44	770	253	232	1255	61.4%	20.2%	18.5%
	3 = 45-64	890	375	278	1543	57.7%	24.3%	18.0%
	4 = 65+	696	339	246	1281	54.3%	26.5%	19.2%
sex	0 = Male	1358	359	210	1927	70.5%	18.6%	10.9%
	1 = Female	1804	789	650	3243	55.6%	24.3%	20.0%
smoker	0 = Nonsmoker	2484	902	625	4011	61.9%	22.5%	15.6%
	1 = Smoker	678	246	235	1159	58.5%	21.2%	20.3%
symptom	0 = no asthma symptom	1768	485	324	2577	68.6%	18.8%	12.6%
	1 = has asthma symptom	1394	663	536	2593	53.8%	25.6%	20.7%

7.3 Data Analysis

To fit a multinomial logit regression model to the asthma data, we need to select a reference category. The NR code created to estimate the parameters for multinomial logistic regression uses the last category as the reference category by default, unless otherwise specified. The multi-categorical response variable of interest Y_{ij} represents the observed counts of the j^{th} category in the i^{th} case. An important feature of the multinomial logit model is that it estimates $c - 1$ models, where $c = 3$ is the number of levels of the outcome/response variable. To describe the effects of the covariates on the multi-categorical response variable, we consider the multi-logistic model below.

$$\log\left(\frac{\pi_{ij}}{\pi_{i3}}\right) = \beta_{0j} + \beta_{1j}age1_{ij} + \beta_{2j}age2_{ij} + \beta_{3j}age3_{ij} \\ + \beta_{4j}sex_{ij} + \beta_{5j}smoker_{ij} + \beta_{6j}symptom_{ij}, \quad (7.1)$$

for $i = 1, \dots, 32$ and $j = 1, 2, 3$. In this instance, the reference category is “Heavy”. So we estimate a model for “Light” relative to “Heavy”, and again a model for “Moderate” relative to “Heavy”. The unrestricted estimates of the parameters in equation (7.1) were obtained using the R function `mnlogit.NR`, and are presented in Table 7.3.

Table 7.3: Unrestricted MLE for multinomial logit of Asthma

Parameter Estimates						
# of consultations		$\hat{\beta}$	Std. Error	Wald	Pr(> Wald)	Exp($\hat{\beta}$)
Light	Intercept	1.920508	0.108145	315.370230	0.000000	6.824422
	age1	0.919048	0.130259	49.780734	0.000000	2.506902
	age2	0.282431	0.108991	6.715013	0.009560	1.326351
	age3	0.255637	0.103987	6.043569	0.013957	1.291284
	sex	-0.739956	0.088663	69.651272	0.000000	0.477135
	smoker	-0.362826	0.091460	15.737624	0.000073	0.695707
	symptom	-0.661094	0.080688	67.128656	0.000000	0.516286
Moderate	Intercept	0.686065	0.123700	30.760202	0.000000	1.985886
	age1	0.220624	0.149836	2.168079	0.140902	1.246855
	age2	-0.162370	0.125757	1.667028	0.196657	0.850127
	age3	0.044762	0.116785	0.146908	0.701508	1.045779
	sex	-0.321443	0.102522	9.830437	0.001716	0.725102
	smoker	-0.308016	0.106795	8.318567	0.003924	0.734904
	symptom	-0.159126	0.093261	2.911289	0.087962	0.852889

Since the parameter estimates are relative to the reference category, the interpretation of the multinomial logit is that for a unit change in the continuous predictor variable, the logit (log-odds) of outcome category j (which is light or moderate) relative to the reference category

(heavy) is expected to change by its respective parameter estimate (which is in log-odds units) given all other model variables are held constant [32]. On the other hand, interpreting the multinomial logit for level m of the categorical predictor relative to the reference category of the same predictor is expected to change by its respective parameter estimate (which is in log-odds units) given all other model variables are held constant.

Symptoms

For example, the odds (log-odds) of asthmatic respondents showing symptoms requiring light health system use rather than heavy is 0.5163 (-0.6611) times lower than the odds of a non-symptomatic asthmatic respondent.

Smoker

The odds (log-odds) of asthmatic respondents, who are also smokers, requiring light health system use rather than heavy is 0.6957 (-0.3628) times lower than the odds of a non-smoker asthmatic respondent.

Sex

The odds (log-odds) of male asthmatic respondents requiring light health system use rather than heavy is 0.4771 (-0.7400) times lower than the odds of female asthmatic respondents.

Age

Another example of this can be seen with the age covariates. For instance, the odds (log-odds) of asthmatic respondents that fall into Age1, Age2, and Age3 that require light health system use rather than heavy are 0.9190 (2.5069), 0.2824 (1.3264), 0.2556 (1.2913) times higher, respectively, than the odds of an asthmatic respondent that falls into Age4.

Similarly to the light relative to the heavy health system use, the odds (log-odds) are observed with the same pattern for moderate relative to heavy health system use for all the categorical predictors (Symptoms, Smoker, Sex, and Age) relative to the reference category of the same predictor, with the exception of Age2 relative to Age4, where the odds are lower rather than higher.

The other columns in Table 7.3 are explained below:

Std. Error: These are the standard errors of the MLEs of the regression coefficients.

Wald: Wald test is used to test the null hypothesis that the individual coefficient is 0.

p-value: the p-value is calculated for testing the null hypothesis that a particular predictor's regression coefficient is zero.

7.4 Restricted Inference

The unrestricted estimates of the multinomial logit parameters in (7.1) are presented in Table 7.4. Here, the unrestricted estimates for β_1 satisfy the constraints; however, the unrestricted estimates for the second set β_2 do not satisfy the constraints under the parameter space $\mathcal{C} = \{\beta = (\beta_1^T, \beta_2^T)^T : \beta_{11} \geq \beta_{21} \geq \beta_{31} \text{ and } \beta_{12} \geq \beta_{22} \geq \beta_{32}\}$, where β_1 and β_2 are 7×1 parameter vectors for light and moderate categories of health system use.

We restrict the age groups to have an increasing effect on the number of medical doctor consultations (i.e. light, moderate vs. heavy use of the health system). More specifically, the parameter space under H_0 is

$$\mathcal{C}_0 = \{\beta = (\beta_1^T, \beta_2^T)^T : \beta_{11} = \beta_{21} = \beta_{31} \text{ and } \beta_{12} = \beta_{22} = \beta_{32}\}, \quad (7.2)$$

while under H_1 is

$$\mathcal{C}_1 = \{\beta = (\beta_1^T, \beta_2^T)^T : \beta_{11} \geq \beta_{21} \geq \beta_{31} \text{ and } \beta_{12} \geq \beta_{22} \geq \beta_{32}\}. \quad (7.3)$$

To obtain the restricted estimates for the multinomial logit model in Eq. (7.1), we used the modified GP algorithm described in Section 6.4. For more details on the GP, see Chapter 4, Theorem 4.2. This algorithm was used to find the constrained estimates for all models as shown in Table 7.4. We can rewrite the parameter space $\mathcal{C}$ in the form of pairwise contrasts $\beta_{\ell j} - \beta_{mj}$. For the light category, the constraints are:

$$\beta_{11} - \beta_{21} \geq 0, \text{ and}$$

$$\beta_{21} - \beta_{31} \geq 0.$$

For the moderate category, the constraints are:

$$\beta_{12} - \beta_{22} \geq 0, \text{ and}$$

$$\beta_{22} - \beta_{32} \geq 0.$$

Therefore, the constraint matrix can be written as:

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (7.4)$$

The hypotheses of interest are:

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b}, \quad H_1 : \mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}, \quad \text{and} \quad H_2 : \text{no restriction.}$$

Let $\boldsymbol{\beta}^0$, $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ be the MLEs under H_0 , H_1 , and H_2 , respectively.

Table 7.4: Unrestricted and Restricted MLE for multinomial logit of Asthma

Parameter Estimates						
# of Consultations	Covariate		Unrestricted		Restricted	
		Parameters	$\hat{\beta}$	Std. Error	β^0	$\tilde{\beta}$
Light	Intercept	β_{01}	1.9205	0.1081	1.982	1.9192
	age1	β_{11}	0.919	0.1303	0.4268	0.9198
	age2	β_{21}	0.2824	0.109	0.4268	0.3474
	age3	β_{31}	0.2556	0.104	0.4268	0.2057
	sex	β_{41}	-0.74	0.0887	-0.7879	-0.7377
	smoker	β_{51}	-0.3628	0.0915	-0.3949	-0.3679
	symptom	β_{61}	-0.6611	0.0807	-0.7063	-0.6604
Moderate	Intercept	β_{02}	0.6861	0.1237	0.7078	0.6837
	age1	β_{12}	0.2206	0.1498	0.0071	0.2219
	age2	β_{22}	-0.1624	0.1258	0.0071	-0.0429
	age3	β_{32}	0.0448	0.1168	0.0071	-0.0429
	sex	β_{42}	-0.3214	0.1025	-0.3331	-0.3174
	smoker	β_{52}	-0.308	0.1068	-0.3271	-0.3167
	symptom	β_{62}	-0.1591	0.0933	-0.1729	-0.1581

The inequality constraints are noted to be true as can be seen from the goodness-of-fit test, where the test statistic $T_{12} = 2[\ell(\hat{\beta}) - \ell(\tilde{\beta})] = 2[-4658.87576 + 4660.34208] = 2.93265$ with p -value = 0.212 computed using Eq. (6.9). Since the alternative hypothesis does not involve equality constraints, there is no need to modify the chi-bar-square weights. For more information about computation of chi-bar-square weights, refer to Theorem 5.4.

The LRT statistic for H_0 against $H_1 - H_0$ is $T_{01} = 2[\ell(\tilde{\beta}) - \ell(\beta^0)] = 2[-4660.34208 + 4682.68126] = 44.67836$ with p -value < 0.00001, computed using Eq. (6.8). We reject H_0 and conclude the increasing age has significant and directional effects on the number of consultations with medical doctors (light, moderate versus heavy health system use), i.e., as age

increases, the probability of having an increased number of consultations with medical doctors also increases. Moreover, the unconstrained test H_0 against $H_2 - H_1$ has LRT statistic $T_{02} = 2[\ell(\hat{\beta}) - \ell(\beta^0)] = 2[-4658.87576 + 4682.68126] = 47.61101$ with p -value < 0.00001 . This suggests that age is significant to health system use; however, no directional effects are indicated. Therefore, the constrained LRT provides additional information which would not otherwise be detected using unconstrained hypothesis tests.

Chapter 8

Constrained Statistical Inference in Multivariate GLMM for Multinomial Data

8.1 Introduction

There are various ways in which to approach modelling clustered data. Since studies often have data that occur in clusters (such as repeated measurements for each subject in a study), we can approach modelling this type of data using random effects for the subjects or clusters in the linear predictor. In this way, correlation structures amongst the clustered observations and overdispersion amongst discrete responses (modelled with Poisson or binomial models) can be accounted for. We commonly see random effects models being used to model heterogeneity and dependence of responses in repeated measurements.

GLMs extend to GLMMs where both random effects and fixed effects are included in the linear predictor when the response variables belong to the exponential family. In GLMMs, the response distribution is defined conditionally on the random effects, which are commonly assumed to be multivariate normal, although this assumption is not necessary. Lee and Nelder [70] use a conjugate distribution as a different approach. The marginal distribution of the response does not have a closed form due to the integration of the normal random effects. To approximate the marginal distribution, the MLE and their standard errors, many numerical integration methods are employed using Gauss-Hermite (GH) quadrature [85] and [86], Monte Carlo techniques [87], [88], [89], and [90], or the Laplace approximation with Taylor series

expansions [91] and [92].

The majority of research to this day remains heavily involved in GLMMs for (conditionally) binomial [93], [86] and [94] and Poisson [85] and [95] distributed responses, with minimal exploration and development in the area of multinomial responses. Even in the case of multinomial responses, the focus has been centred around ordinal models with logit and probit link functions for cumulative probabilities. Various researchers contributed different models, each with its own techniques and outcomes: Harville and Mee [96] proposed a cumulative probit random effects model, which relied on Taylor series approximations for intractable integrals; Jansen [97] and Ezzet and Whitehead [98] proposed random intercept cumulative probit and logit models, respectively, each of whom employed quadrature techniques. More general ordinal random effects models allowing for multiple random effects were proposed by Hedeker and Gibbons [99]; they applied Gauss-Hermite quadrature within a Fisher scoring algorithm. Tutz and Hennevogl [73] used quadrature and Monte Carlo EM algorithms.

Modelling baseline categorical logit for multinomial responses with random effects has not received much attention. Few areas, such as economics, epidemiology and psychology make use of this modelling technique and its statistical applications. This type of modelling can be applied when subjects respond to a number of related multiple choice questions where the response categories are unordered.

This chapter presents a general approach for logit random effects modelling of clustered nominal responses. We review multinomial logit random effects models in a unified form as multivariate generalized linear mixed models (MGLMM). We often see the use of multinomial logit models with random effects in biomedical science when analysing correlated nominal data, which can be a result of repeated measures or clustered observations. This type of computation, however, can be costly due to the presence of random effects and the complexity of the likelihood function; the presence of random effects brings with it the computationally expensive multi-dimensional integrals since the number of points for integral computation increases exponentially. For the purpose of this thesis, we calculate the MLE that uses Adaptive Gauss-Hermite Quadrature

(AGQ) maximization algorithm. This chapter also covers constrained inference for MGLMM using GP algorithm.

To address costly computations, normally, the stochastic Monte Carlo EM algorithm is used. Hartzel et. al [76] attempts to generalize a pseudo-likelihood approach that is simpler; however, it provides poorer approximations for the likelihood. A number of approaches have also been proposed by numerous authors over the course of this field's development, including a penalized quasi-likelihood approach that avoids the complex form of the multinomial likelihood [91], where we transform the multinomial problem to a Poisson log-linear or non-linear model [77] and [42], and the use of Lagrange multipliers to normalize multinomial probabilities while noting that due to enforcing normalization for each distinct covariate pattern, the Poisson log-linear transformation is restricted to discrete covariates [100].

8.2 Random Effects Models for Nominal Data

Here random effects models for binary responses extend to multicategory responses. For the multicategory models of Chapter 3, a multinomial observation with c categories is explained. In Section 3.4, we defined a multivariate GLM by applying a vector of link function to this multivariate response. Adding random effects extends this multivariate GLM to a multivariate GLMM [76] [73]. We will look at models for nominal responses.

8.2.1 Model Specification

Suppose cluster i has n_i categorical observations. Let Y_{it} denote the t^{th} observation in cluster $i = 1, \dots, m$, $t = 1, \dots, n_i$, with response probabilities $\pi_{itj} = P(Y_{it} = j | \mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{u}_i)$. Let $\mathbf{x}_{it}$ denote a $(p+1)$ -dimensional column vector of explanatory variables for that observation where $x_{it0} = 1$, and let $\mathbf{z}_{it}$ denote a $v = (s+1)$ -dimensional column vector of coefficients for the random effects $z_{it0} = 1$ [76]. Let g_j and $h_j = g_j^{-1}$ be the link and inverse link functions, respectively. The random-effects model assumes that $\mathbf{u}_i$ are mutually independent from a $N(\mathbf{0}, \mathbf{\Sigma})$ distribution,

with Σ unknown, and that conditionally on the $\mathbf{u}_i$ the multivariate response vector for the t^{th} observation in cluster i , $\mathbf{y}_{it}$, are independent. Unconditionally, variability among $\mathbf{u}_i$ reflects cluster heterogeneity, whereby different clusters have different response probabilities [71].

8.2.2 Baseline-Category Logit Models with Random Effects

When the categories for Y_{it} are unordered, referring to the nominal response variables, we use the logit model to pair each response category with an arbitrary baseline category (for example, last category c), and fit these models simultaneously. The general form of the baseline-category logit model with random effects is given by

$$\eta_{itj} = g_j(\pi_{itj}) = \text{logit}(\pi_{itj}) = \log \left(\frac{P(Y_{it} = j | \mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{u}_i)}{P(Y_{it} = c | \mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{u}_i)} \right) = \mathbf{x}_{it}^T \boldsymbol{\beta}_j + \mathbf{z}_{it}^T \mathbf{u}_i, \quad (8.1)$$

$$\pi_{itj} = h(\eta_{itj}) = \frac{\exp(\mathbf{x}_{it}^T \boldsymbol{\beta}_j + \mathbf{z}_{it}^T \mathbf{u}_i)}{1 + \sum_{j=1}^g \exp(\mathbf{x}_{it}^T \boldsymbol{\beta}_j + \mathbf{z}_{it}^T \mathbf{u}_i)}, \quad (8.2)$$

where $\boldsymbol{\beta}_j = (\beta_{0j}, \dots, \beta_{pj})^T$ is a fixed coefficient vector and $\mathbf{u}_i = (u_{i0}, \dots, u_{is})^T$ is a $(s+1)$ -dimensional cluster-specific random effect. The models (8.1) and (8.2) can be presented in matrix form as multivariate generalized linear mixed models (MGLMMs) for categorical responses. We use the notation Y_{it} for the t^{th} observation of cluster i . Y_{it} takes values from $\{1, \dots, c\}$, or $\mathbf{y}_{it} = (y_{it1}, \dots, y_{itg})^T$. We summarize the observations as $(\mathbf{y}_{it}^T, \mathbf{x}_{it}^T, \mathbf{z}_{it}^T)$, $i = 1, \dots, m$, $t = 1, \dots, n_i$. The corresponding model for observation $\mathbf{y}_{it}$ has the form:

$$\boldsymbol{\eta}_{it} = \mathbf{g}(\boldsymbol{\pi}_{it}) = \text{logit}(\boldsymbol{\pi}_{it}) = \mathbf{X}_{it} \boldsymbol{\beta} + \mathbf{Z}_{it} \mathbf{u}_i, \quad (8.3)$$

$$\boldsymbol{\pi}_{it} = \mathbf{h}(\boldsymbol{\eta}_{it}), \quad (8.4)$$

where the linear predictor $\boldsymbol{\eta}_{it} = (\eta_{it1}, \dots, \eta_{itg})^T$ is the g -dimensional row vector of different logits, $\boldsymbol{\pi}_{it} = (\pi_{it1}, \dots, \pi_{itg})^T$ is the g -dimensional row vector of probabilities, $q = g(p+1)$ -dimensional column vector $\boldsymbol{\beta}$ is the vector for fixed parameters, and $\mathbf{u}_i$ is the vector for the random effects; $g \times q$ -dimensional $\mathbf{X}_{it}$ and the $g \times v$ -dimensional $\mathbf{Z}_{it}$ are the model matrices for the fixed and random effects respectively, all typically have the forms:

$$\mathbf{X}_{it} = \begin{pmatrix} \mathbf{x}_{it}^T & \mathbf{0}^T & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{x}_{it}^T & \cdots & \mathbf{0}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \mathbf{x}_{it}^T \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_g \end{pmatrix}, \quad \mathbf{Z}_{it} = \begin{pmatrix} \mathbf{z}_{it}^T \\ \vdots \\ \mathbf{z}_{it}^T \end{pmatrix}.$$

$\mathbf{u}_i$ usually follows a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The vector of probabilities $\boldsymbol{\pi}_{it} = \mathbb{E}(\mathbf{y}_{it} | \mathbf{u}_i)$ is the conditional mean of $\mathbf{y}_{it}$ on the random effect $\mathbf{u}_i$, $\mathbf{g} = (g_1, \dots, g_g)^T$ is the multivariate link function, and $\mathbf{h} = \mathbf{g}^{-1}$ is the inverse link function. The form of $\mathbf{Z}_{it}$ depends on the structure of the cluster-specific effects. We combine observations from one cluster to obtain:

$$\boldsymbol{\eta}_i = \mathbf{g}(\boldsymbol{\pi}_i) = \text{logit}(\boldsymbol{\pi}_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i, \quad (8.5)$$

$$\boldsymbol{\pi}_i = \mathbf{h}(\boldsymbol{\eta}_i), \quad (8.6)$$

where $\boldsymbol{\eta}_i = (\boldsymbol{\eta}_{i1}^T, \dots, \boldsymbol{\eta}_{in_i}^T)^T$ is the $n_i g$ -dimensional row vector of different logits for cluster i , $\boldsymbol{\pi}_i = (\boldsymbol{\pi}_{i1}^T, \dots, \boldsymbol{\pi}_{in_i}^T)^T$ is the $n_i g$ -dimensional row vector of response probabilities for cluster i , $q \times n_i g$ -dimensional $\mathbf{X}_i^T = (\mathbf{X}_{i1}^T, \dots, \mathbf{X}_{in_i}^T)$, and $v \times n_i g$ -dimensional $\mathbf{Z}_i^T = (\mathbf{Z}_{i1}^T, \dots, \mathbf{Z}_{in_i}^T)$ are the model matrices for the fixed and random effects for cluster i , respectively. Let $N = \sum_{i=1}^m n_i$, for all observations. We can write in matrix form

$$\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\pi}) = \text{logit}(\boldsymbol{\pi}) = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u}, \quad (8.7)$$

$$\boldsymbol{\pi} = \mathbf{h}(\boldsymbol{\eta}), \quad (8.8)$$

where the linear predictor $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^T, \dots, \boldsymbol{\eta}_m^T)^T$ is the Ng -dimensional row vector of different logits, $\boldsymbol{\pi} = (\boldsymbol{\pi}_1^T, \dots, \boldsymbol{\pi}_m^T)^T$ is the Ng -dimensional row vector of probabilities, $q \times Ng$ -dimensional $\mathbf{X}^T = (\mathbf{X}_1^T, \dots, \mathbf{X}_m^T)$ and $Ng \times mv$ -dimensional $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \dots, \mathbf{Z}_m)$ are the model matrices for the fixed and random effects, respectively. We assume a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{u}} = \text{diag}(\boldsymbol{\Sigma}, \dots, \boldsymbol{\Sigma})$ for the mv -dimensional row vector of random effects $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_m^T)^T$.

8.3 Multivariate Likelihood Function

Let $\mathbf{y}_{it}^T | \mathbf{u}_i = (y_{it1}, \dots, y_{itg}) \sim MN(n_{it}, \boldsymbol{\pi}_{it})$, $i = 1, \dots, m$, $t = 1, \dots, n_i$, denote the multinomial distribution with $c = g + 1$ categories. The multinomial distribution has the form of a multivariate exponential family (see Appendix B). This means that the multinomial logit random effects models are a special case of MGLMM. The conditional density of $\mathbf{y}_{it}$, given the explanatory variables, $\mathbf{X}_{it}$ and $\mathbf{Z}_{it}$ in Eq. (8.3), and the v -dimensional random effect $\mathbf{u}_i$, $f(\mathbf{y}_{it} | \mathbf{u}_i)$ belongs to the multivariate exponential family with

$$\boldsymbol{\mu}_{it} = E(\mathbf{y}_{it} | \mathbf{u}_i) = \mathbf{h}(\boldsymbol{\eta}_{it}), \quad \boldsymbol{\eta}_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{Z}_{it}\mathbf{u}_i. \quad (8.9)$$

The general multinomial model is defined by equation (8.9) in terms of the response vector $\mathbf{y}_{it}$ or scaled multinomials/proportions $\mathbf{p}_{it} = \frac{1}{n_{it}}\mathbf{y}_{it}$. For example, the baseline-category logit random effects model has

$$\pi_{itj} = h_j(\boldsymbol{\eta}_{it}) = \frac{\exp(\eta_{itj})}{1 + \sum_{j=1}^g \exp(\eta_{itj})}, \quad j = 1, \dots, g \quad \text{and} \quad \pi_{itc} = \frac{1}{1 + \sum_{j=1}^g \exp(\eta_{itj})}.$$

Then the conditional probability mass function is

$$\begin{aligned} f(\mathbf{y}_{it} | \mathbf{u}_i) &= \frac{n_{it}!}{\prod_{j=1}^c y_{itj}!} \prod_{j=1}^c \pi_{itj}^{y_{itj}} \\ &= \frac{n_{it}!}{y_{it1}! \dots y_{itg}! (n_{it} - \sum_{j=1}^g y_{itj})!} \pi_{it1}^{y_{it1}} \dots \pi_{itg}^{y_{itg}} (1 - \sum_{j=1}^g \pi_{itj})^{(n_{it} - \sum_{j=1}^g y_{itj})} \\ &= \exp \{ \mathbf{y}_{it}^T \boldsymbol{\theta}_{it} + n_{it} \log(\pi_{itc}) + \log(c_{it}) \} \\ &= \exp \{ n_{it} \mathbf{p}_{it}^T \boldsymbol{\theta}_{it} + n_{it} \log(\pi_{itc}) + \log(c_{it}) \}, \end{aligned} \quad (8.10)$$

where the canonical parameter vector is $\boldsymbol{\theta}_{it} = (\theta_{it1}, \dots, \theta_{itg})^T$, $\theta_{itj} = \log\left(\frac{\pi_{itj}}{\pi_{itc}}\right)$, $\pi_{itc} = 1 - \sum_{j=1}^g \pi_{itj}$, and the dispersion parameter is $\frac{1}{n_{it}}$, $c_{it} = \frac{n_{it}!}{\prod_{j=1}^c y_{itj}!}$. Averaging out the continuous random effect through integration, the marginal distribution has mean

$$\mathbf{E}(\mathbf{y}_{it}) = \mathbf{E}[\mathbf{E}(\mathbf{y}_{it}|\mathbf{u}_i)] = \mathbf{E}[\mathbf{h}(\boldsymbol{\eta}_{it})],$$

and variance-covariance matrix

$$\mathbf{V}(\mathbf{y}_{it}) = \mathbf{E}[\mathbf{V}(\mathbf{y}_{it}|\mathbf{u}_i)] + \mathbf{V}[\mathbf{E}(\mathbf{y}_{it}|\mathbf{u}_i)].$$

The distribution of the total response for the i^{th} cluster $n_{ig} \times 1$ -vector $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \dots, \mathbf{y}_{in_i}^T)^T = (y_{i11}, \dots, y_{i1g}, \dots, y_{in_i1}, \dots, y_{in_ig})^T$ is obtained by assuming the conditional independence of $\mathbf{y}_{i1}, \dots, \mathbf{y}_{in_i}$ given $\mathbf{u}_i$. The marginal probability function of $\mathbf{y}_i$ is

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i, \mathbf{u}_i) d\mathbf{u}_i = \int f(\mathbf{y}_i|\mathbf{u}_i) \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i = \int \left[\prod_{t=1}^{n_i} f(\mathbf{y}_{it}|\mathbf{u}_i) \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i, \quad (8.11)$$

where $\phi(\mathbf{u}_i, \boldsymbol{\Sigma})$ denotes the density of the random effects, which are assumed to have no relations with the fixed effects. The expression (8.11) always involves intractable integrals whose dimension depends on the structure of the random effects. In general, the GLMM likelihood function is the marginal mass function of the observed multinomial data, $\mathbf{y}$, viewed as a function of the parameters, and has the form:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \prod_{i=1}^m f(\mathbf{y}_i) = \prod_{i=1}^m \int \left[\prod_{t=1}^{n_i} f(\mathbf{y}_{it}|\mathbf{u}_i) \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i \\ &= \prod_{i=1}^m \int \left[\prod_{t=1}^{n_i} \exp \{ \mathbf{y}_{it}^T \boldsymbol{\theta}_{it} + n_{it} \log(\pi_{itc}) + \log(c_{it}) \} \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i. \end{aligned} \quad (8.12)$$

Note that the fixed effects $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$ are the unknown parameters to be estimated, where the covariance matrix $\boldsymbol{\Sigma}$ of the random effects $\mathbf{u}_i$ depends on an unknown parameter vector $\boldsymbol{\sigma}$, which represents the variance components. Estimates of random effects models can be obtained in various ways by solving integrals numerically, either by the deterministic methods (Gaussian quadrature or AGQ) or the stochastic methods (Monte Carlo Markov Chain). The first method uses AGQ when the dimension of the random effect is low, where AGQ is the best method for approximating the multivariate normal integrals, which is basically a discrete

approximation of the multivariate normal integral, see Definition 8.1. An alternative approach that can be extended to MGLMM is penalized quasi-likelihood, which has been suggested by Breslow and Clayton [91]. A second method that can be used is the Monte Carlo method to simulate the likelihood rather than computing it directly. Using Eugene Demidenko's text as a reference [41], we note the following definition:

Definition 8.1 (Deterministic Method): A proper integral (over a finite interval) can be approximated by a finite weighted sum with any predefined precision $\epsilon > 0$,

$$\int_{-\infty}^{\infty} f(x)dx \approx \int_A^B f(x)dx \approx \sum_{k=1}^K w_k f(x_k),$$

where $A = x_1 < x_2 < \dots < x_K = B$ are the nodes (knots or abscissas) and w_k are the positive weights. A and B are called the lower and upper limits of the integration, and careful attention must be made when choosing them.

In the Gauss-Hermite (GH) Quadrature, nodes (abscissas) and weights are available when function f is proportional to e^{-x^2} . The closer f is to e^{-x^2} , the better the precision of the GH quadrature. Abscissas and weights, $\{x_k, w_k, k = 1, \dots, K\}$, up to $K = 20$, are a part of GH quadrature for evaluating an integral of the form $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$. After (x_k, w_k) are determined, the integral is approximated as a simple sum,

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx \approx \sum_{k=1}^K w_k f(x_k).$$

The GH quadrature for two or three-dimensional integrals with the Gaussian kernel are expressed as:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)e^{-x^2-y^2}dx \approx \sum_{k=1}^K \sum_{k'=1}^K w_k w_{k'} f(x_k, y_{k'}),$$

and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z)e^{-x^2-y^2-z^2}dx \approx \sum_{k=1}^K \sum_{k'=1}^K \sum_{k''=1}^K w_k w_{k'} w_{k''} f(x_k, y_{k'}, z_{k''}), \text{ respectively.}$$

It is straightforward to generalize the GH quadrature to multidimensional integrals.

8.3.1 Maximum Likelihood

The MGLMM with multiple linear random effects takes the form $\boldsymbol{\eta}_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{Z}_{it}\mathbf{u}_i$, where $\mathbf{u}_i$ is the v -dimensional random effect. In this section, we assume that the random effects are independent normally distributed, $\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ is subject to estimation along with the $\boldsymbol{\beta}$ coefficients. The advantage of the normally distributed random effects is that $\mathbf{Z}_{it}\mathbf{u}_i$ is also normally distributed as $N(\mathbf{0}, \mathbf{Z}_{it}\boldsymbol{\Sigma}\mathbf{Z}_{it}^T)$. A vector-valued random variable $\mathbf{u}_i$ is said to have a multivariate normal (or Gaussian) distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ if its probability density function is given by

$$\phi(\mathbf{u}_i, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^v |\boldsymbol{\Sigma}|}} \exp\{-0.5\mathbf{u}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_i\}. \quad (8.13)$$

Expressing the likelihood function in terms of the precision matrix $\boldsymbol{\Sigma}$, we have

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \prod_{i=1}^m f(\mathbf{y}_i) = \prod_{i=1}^m \int \left[\prod_{t=1}^{n_i} f(\mathbf{y}_{it} | \mathbf{u}_i) \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i \\ &= \prod_{i=1}^m \int \left[\prod_{t=1}^{n_i} c_{it} \left(\frac{\pi_{it1}}{\pi_{itc}} \right)^{y_{it1}} \cdots \left(\frac{\pi_{itg}}{\pi_{itc}} \right)^{y_{itg}} \pi_{itc}^{n_{it}} \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i \\ &= \prod_{i=1}^m \int \left[\prod_{t=1}^{n_i} c_{it} e^{y_{it1}\eta_{it1}} \cdots e^{y_{itg}\eta_{itg}} \left(1 + \sum_{j=1}^g e^{\eta_{itj}} \right)^{-n_{it}} \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i \\ &= \prod_{i=1}^m \int \left[C_i \exp \left\{ \sum_{t=1}^{n_i} \sum_{j=1}^g y_{itj} \eta_{itj} \right\} \prod_{t=1}^{n_i} \left(1 + \sum_{j=1}^g e^{\eta_{itj}} \right)^{-n_{it}} \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i \\ &= \prod_{i=1}^m \int \left[C_i \exp \left\{ \sum_{t=1}^{n_i} \sum_{j=1}^g y_{itj} (\mathbf{x}_{it}^T \boldsymbol{\beta}_j + \mathbf{z}_{it}^T \mathbf{u}_{ij}) \right\} \prod_{t=1}^{n_i} \left(1 + \sum_{j=1}^g e^{\eta_{itj}} \right)^{-n_{it}} \right] \phi(\mathbf{u}_i, \boldsymbol{\Sigma}) d\mathbf{u}_i \\ &= \prod_{i=1}^m \frac{C_i}{\sqrt{(2\pi)^v |\boldsymbol{\Sigma}|}} \exp \left\{ \sum_{t=1}^{n_i} \sum_{j=1}^g y_{itj} \mathbf{x}_{it}^T \boldsymbol{\beta}_j \right\} \times \\ &\quad \int \exp \left\{ \left(\sum_{t=1}^{n_i} \sum_{j=1}^g y_{itj} \mathbf{z}_{it}^T \mathbf{u}_{ij} \right) - 0.5 \mathbf{u}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_i - \sum_{t=1}^{n_i} n_{it} \ln \left(1 + \sum_{j=1}^g e^{\eta_{itj}} \right) \right\} d\mathbf{u}_i \\ &= \prod_{i=1}^m \frac{C_i}{\sqrt{(2\pi)^v |\boldsymbol{\Sigma}|}} e^{\mathbf{r}_i^T \boldsymbol{\beta}} \int \exp \left\{ \mathbf{k}_i^T \mathbf{u}_i - 0.5 \mathbf{u}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_i - \sum_{t=1}^{n_i} n_{it} \ln \left(1 + \sum_{j=1}^g e^{\eta_{itj}} \right) \right\} d\mathbf{u}_i, \end{aligned} \quad (8.14)$$

where $C_i = \prod_{t=1}^{n_i} c_{it} = \prod_{t=1}^{n_i} \frac{n_{it}}{y_{it1}! \cdots y_{itg}!}$, $\mathbf{r}_i^T = \sum_{t=1}^{n_i} \mathbf{y}_{it}^T \mathbf{X}_{it} = \sum_{t=1}^{n_i} (y_{it1} \mathbf{x}_{it}^T, \dots, y_{itg} \mathbf{x}_{it}^T)$, and

$$\mathbf{k}_i^T = \sum_{t=1}^{n_i} \mathbf{y}_{it}^T \mathbf{Z}_{it} = \sum_{t=1}^{n_i} (y_{it1} \mathbf{z}_{it}^T, \dots, y_{itg} \mathbf{z}_{it}^T).$$

In addition, $\mathbf{r}_i^T$ and $\mathbf{k}_i^T$ can be computed as $\text{vec}(\mathbf{X}_i^T \mathbf{Y}_i)$ and $\text{vec}(\mathbf{Z}_i^T \mathbf{Y}_i)$, respectively, where $\mathbf{Y}_i$ is a $n_i \times g$ response matrix. The likelihood function can be rewritten as

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \frac{C}{\sqrt{(2\pi)^{mv} \prod_{i=1}^m (|\boldsymbol{\Sigma}|)}} \exp \{ \mathbf{r}^T \boldsymbol{\beta} \} \prod_{i=1}^m \int e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i, \quad (8.15)$$

where

$$h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathbf{k}_i^T \mathbf{u}_i - 0.5 \mathbf{u}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_i - \sum_{t=1}^{n_i} n_{it} \ln \left(1 + \sum_{j=1}^g e^{\eta_{itj}} \right),$$

$$\mathbf{r}^T = \sum_{i=1}^m \mathbf{r}_i^T, \quad C = \prod_{i=1}^m C_i \quad \text{and} \quad \eta_{itj} = \mathbf{x}_{it}^T \boldsymbol{\beta}_j + \mathbf{z}_{it}^T \mathbf{u}_i.$$

Therefore the kernel of the log-likelihood function takes the form

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \ln L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &= -0.5mv \ln(2\pi) - 0.5 \sum_{i=1}^m \ln(|\boldsymbol{\Sigma}|) + \mathbf{r}^T \boldsymbol{\beta} + \sum_{i=1}^m \ln \left(\int e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i \right). \end{aligned} \quad (8.16)$$

The MLEs for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^{-1}$ are the solution to the score equations, $\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\beta}} = \mathbf{0}$ and $\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} = \mathbf{0}$.

The first-order derivatives are:

$$\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\beta}} = \mathbf{r}^T + \sum_{i=1}^m \frac{\mathbf{I}_{i3}}{I_{i1}}, \quad \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2} \left(m(\boldsymbol{\Sigma}^{-1})^{-1} - \sum_{i=1}^m \frac{\mathbf{I}_{i2}}{I_{i1}} \right), \quad (8.17)$$

where

$$I_{i1} = \int e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i, \quad \mathbf{I}_{i2} = \int \mathbf{u}_i \mathbf{u}_i^T e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i, \quad \text{and}$$

using Leibniz rule:

$$\begin{aligned}
\mathbf{I}_{i3} &= \frac{\partial}{\partial \boldsymbol{\beta}} \int e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i = \int \frac{\partial e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})}}{\partial \boldsymbol{\beta}} d\mathbf{u}_i = \int \frac{\partial h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\beta}} e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i \\
&= \int \begin{pmatrix} \frac{\partial h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \beta_1} \\ \vdots \\ \frac{\partial h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \beta_g} \end{pmatrix} e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i = - \int \begin{pmatrix} \sum_{t=1}^{n_i} n_{it} \frac{e^{\eta_{it1}}}{1 + \sum_{j=1}^g e^{\eta_{itj}}} \mathbf{x}_{it}^T \\ \vdots \\ \sum_{t=1}^{n_i} n_{it} \frac{e^{\eta_{itg}}}{1 + \sum_{j=1}^g e^{\eta_{itj}}} \mathbf{x}_{it}^T \end{pmatrix} e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i \\
&= - \int \begin{pmatrix} \sum_{t=1}^{n_i} n_{it} \pi_{it1} \mathbf{x}_{it}^T \\ \vdots \\ \sum_{t=1}^{n_i} n_{it} \pi_{itg} \mathbf{x}_{it}^T \end{pmatrix} e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i = - \int \begin{pmatrix} \sum_{t=1}^{n_i} \mu_{it1} \mathbf{x}_{it}^T \\ \vdots \\ \sum_{t=1}^{n_i} \mu_{itg} \mathbf{x}_{it}^T \end{pmatrix} e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i \\
&= - \int \left[\sum_{t=1}^{n_i} n_{it} \boldsymbol{\pi}_{it}^T \mathbf{X}_{it} \right] e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i = - \int \mathbf{a}_i^T e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i.
\end{aligned}$$

The score equations may be solved iteratively by the Empirical Fisher Scoring (EFS) algorithm since the derivative $\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}}$ is easy to compute. We can also apply the Fixed Point algorithm for the precision matrix, $\boldsymbol{\Sigma}^{-1} = m \left(\sum_{i=1}^m \mathbf{I}_{i2} I_i^{-1} \right)^{-1}$. Although a mixed model rarely has a large number of random effects (typically, $v = 2$ or 3), multidimensionality may substantially increase computation time, which is problematic. Moreover, improper integrals also pose a difficult problem of numerical quadrature.

This thesis will focus only on the random intercept model to avoid the complexities described above in relation to numerical integration for multi-dimensional random effects for each of the g categories. Section 8.4 will cover the random intercept multinomial logit model.

8.4 Random Intercept Multinomial Logit Model

In order to address cluster heterogeneity and intracluster correlation we use the random intercept multinomial model that is a special case of MGLMM. We will look at the simple model containing only one random intercept. Generally speaking, in random intercept models, the linear predictor, η_{itj} , of observations in the i^{th} cluster for the j^{th} category is given by:

$$\eta_{itj} = g_j(\pi_{itj}) = \text{logit}(\pi_{itj}) = \log \left(\frac{P(Y_{it} = j | \mathbf{x}_{it}, u_i)}{P(Y_{it} = c | \mathbf{x}_{it}, u_i)} \right) = \mathbf{x}_{it}^T \boldsymbol{\beta}_j + u_i, \quad (8.18)$$

$$\pi_{itj} = h(\eta_{itj}) = \frac{\exp(\mathbf{x}_{it}^T \boldsymbol{\beta}_j + u_i)}{1 + \sum_{j=1}^g \exp(\mathbf{x}_{it}^T \boldsymbol{\beta}_j + u_i)}, \quad (8.19)$$

where u_i is the cluster-specific intercept for all categories. The fixed effects determine the effects of the covariates but the response strength may vary across different clusters. The random intercept model for clustered data can be obtained by specifying $\mathbf{z}_{it} = \mathbf{1}$ from the general form (8.1). When our data is sparse, i.e., when the number of observations per cluster n_i is small, we use conditional likelihood. For additional information on this specific topic, the reader is referred to Demidenko's work, specifically, Section 7.6 [41]. The model for observation $\mathbf{y}_{it}^T$ has the form:

$$\boldsymbol{\eta}_{it} = \mathbf{X}_{it} \boldsymbol{\beta} + \mathbf{Z}_{it} u_i \quad (8.20)$$

$$\boldsymbol{\pi}_{it} = \mathbf{h}(\boldsymbol{\eta}_{it}), \quad (8.21)$$

where $\mathbf{Z}_{it} = \mathbf{1}_g$ and the intercept random effects u_i , $i = 1, \dots, m$ are assumed to follow an independent normal distribution with the mean 0 and variance component σ^2 as follows

$$u_i \sim \phi_u(0, \sigma^2), \quad (8.22)$$

As in Eq. (8.11), $\phi_u(0, \sigma^2)$ is assumed to have no relation with the fixed effects. From models (8.20) and (8.22), the marginal log-likelihood of $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \sigma^2)^T$, given data $\mathbf{y}$, is expressed as (8.16), where covariance matrix $\boldsymbol{\Sigma}$ is replaced by the variance component σ^2 .

$$\ell(\boldsymbol{\beta}, \sigma^2) = -0.5mv \ln(2\pi) - 0.5m \ln(\sigma^2) + \mathbf{r}^T \boldsymbol{\beta} + \sum_{i=1}^m \ln \left(\int e^{h_i(\mathbf{u}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})} d\mathbf{u}_i \right). \quad (8.23)$$

Then the observed fisher information, $\mathcal{I}_0(\boldsymbol{\delta})$, is decomposed into the form:

$$\mathcal{I}_0(\boldsymbol{\delta}) = \begin{pmatrix} \mathcal{I}_0(\boldsymbol{\beta}, \boldsymbol{\beta}) & \mathcal{I}_0(\boldsymbol{\beta}, \sigma^2) \\ \mathcal{I}_0(\sigma^2, \boldsymbol{\beta}) & \mathcal{I}_0(\sigma^2, \sigma^2) \end{pmatrix}. \quad (8.24)$$

Let $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2)^T$ be the unconstrained ML estimators of $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \sigma^2)^T$. The ML estimator is consistent when the number of clusters, m , goes to infinity and the number of observations per cluster, n_i , is finite. What amplifies the need for ML despite its required integration is that the bulk of approximation methods (such as Laplace or quasi-likelihood) require both m and n_i to go to infinity. Another important argument to favour MLE is that it produces the asymptotic covariance matrix as the inverse of the Fisher information matrix. To obtain the MLE, we can also use the Newton-Raphson, Fisher Scoring, or Expectation-Maximization method.

8.4.1 Unconstrained ML Inference for CCHS Data

In this thesis, the multinomial random intercept logit model for MGLMM was considered for analysing the CCHS data as described in Chapter 7. The health regions will represent the clusters with a total of 97 clusters. Regions are identified as the variable GEODPMF. As mentioned previously, we fit a multinomial random intercept logit model to the asthma data with only one random effect for both categories defined below. To obtain the estimates of the parameters in Eq. (8.25), the built-in R function, *optim*, is used to optimize the likelihood function (8.23), where σ^2 is a single variance component. The multi-categorical response variable of interest Y_{itj} represents the outcomes for the t^{th} observation in cluster i for category j .

$$Y_{itj} = \begin{cases} 1, & \text{if outcome is in } j \text{ category} \\ 0, & \text{otherwise.} \end{cases}$$

An important feature of the multinomial logit model is that it uses $c - 1$ models, where $c = 3$ is the number of levels of the outcome/response variable. To describe the effects of the covariates on the multi-categorical response variable, we consider the multi-logistic random intercept

model below.

$$\begin{aligned} \eta_{itj} = \log\left(\frac{\pi_{itj}}{\pi_{it3}}\right) = & \beta_{0j} + \beta_{1j}age1_{itj} + \beta_{2j}age2_{itj} + \beta_{3j}age3_{itj} \\ & + \beta_{4j}sex_{itj} + \beta_{5j}smoker_{itj} + \beta_{6j}symptom_{itj} + u_i, \end{aligned} \quad (8.25)$$

for $i = 1, \dots, 97$, $t = 1, \dots, n_i$ and $j = 1, 2$. In this instance, the reference category is “Heavy”; and so, we estimate a model for “Light” relative to “Heavy”, and again a model for “Moderate” relative to “Heavy”. The unrestricted estimates of the parameters in model (8.25) were obtained using the *optim* R function, and are presented in Table 8.1.

Table 8.1: Unrestricted MLE for multinomial random intercept logit of Asthma and Odds

Parameter Estimates						
# of consultations		$\hat{\beta}$	Std. Error	Wald	$\Pr(> \text{Wald})$	$\text{Exp}(\hat{\beta})$
Light	Intercept	1.940467	0.113129	294.212827	0.000000	6.962002
	age1	0.943946	0.131567	51.475191	0.000000	2.570104
	age2	0.293828	0.110188	7.110838	0.007700	1.341553
	age3	0.254779	0.104897	5.899306	0.015100	1.290177
	sex	-0.747402	0.089372	69.936680	0.000000	0.473595
	smoker	-0.389032	0.092716	17.606122	0.000000	0.677712
	symptom	-0.664742	0.081414	66.666338	0.000000	0.514406
Moderate	Intercept	0.706011	0.128062	30.393448	0.000000	2.025895
	age1	0.245604	0.150985	2.646059	0.103800	1.278393
	age2	-0.150884	0.126803	1.415884	0.234100	0.859948
	age3	0.044138	0.117595	0.140880	0.707400	1.045127
	sex	-0.328923	0.103149	10.168625	0.001400	0.719699
	smoker	-0.334780	0.107929	9.621580	0.001900	0.715495
	symptom	-0.162685	0.093901	3.001650	0.083200	0.849858
Variance	σ^2	0.074528	0.031609			

We note that the variance component of the random effect, **0.074528**, is very small; and so,

we can deduce that there is little to no difference between the fixed coefficient estimate from the multinomial logit models and the multinomial random intercept logit model. However, the multinomial random intercept logit model provides additional information about the regional random effect, where its variance components appear to be significant. It is worthwhile to note from the p -values that the effects of the covariates (age, sex, smoker, symptoms) are more significant in the multinomial random intercept logit model with the exception of Age3 in both light and moderate vs. heavy categories and of Age2 in the moderate category than those in the multinomial logit model with only fixed effects parameters, as shown in Table 7.3.

8.4.2 Previous Research

Literature is thin when developing algorithms for the MLE under inequality constraints. This thesis' work is based on Jamshidian's [11] use of the GP algorithm for equality and inequality constraints under a general likelihood function. Given the general nature of the GP algorithm's application, its results can be extended to nonlinear mixed models, GLMM and MGLMM.

8.5 Constrained ML Inference for MGLMMs

Inequality constraints in MGLMM are the most challenging problem to tackle. We use the GP algorithm to address this problem. In the following sections, we specify the MGLMM for clustered or longitudinal data, and derive constrained MLEs and LRTs. In this section, we will review restricted ML estimation in MGLMMs using the gradient projection method.

8.5.1 Gradient Projection Algorithm for MGLMMs

Let the convex cone $\mathcal{C}_1 = \{\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T : \mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}\}$ denote the constrained parameter space, where $\mathbf{A}$ is a $r \times p$ matrix of full row rank, $r < p$, $\mathbf{b}$ is $r \times 1$ vector, and where p is the dimension of $\boldsymbol{\beta}$. Here $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \dots, \boldsymbol{\beta}_g^T)^T$, where $\boldsymbol{\beta}_j$ are the model parameters for the j^{th} category. In order to maximize the multinomial log-likelihood function in Eq. (8.16) under such equality and

inequality constraints, we implement a modified version of the gradient projection algorithm of Jamshidian (2004) (for more information see Theorem 4.2), which searches active constraint sets to determine the optimal solution, and another modified GP algorithm referenced in Section 6.4. In the MGLMM, unlike Jamshidian (2004), we maximize the marginal log-likelihood $\ell(\boldsymbol{\delta}|\mathbf{y})$ under linear inequality constraints on the regression parameters. Since the parameter vector is composed of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$, we partition the inverse of the observed information matrix (8.24) as follows:

$$\mathcal{I}^{-1}(\boldsymbol{\delta}) = \mathbf{W}(\boldsymbol{\delta}) = \begin{pmatrix} \mathbf{W}_{11}(\boldsymbol{\delta}) & \mathbf{W}_{12}(\boldsymbol{\delta}) \\ \mathbf{W}_{21}(\boldsymbol{\delta}) & \mathbf{W}_{22}(\boldsymbol{\delta}) \end{pmatrix}. \quad (8.26)$$

where

$$\begin{aligned} \mathbf{W}_{11}(\boldsymbol{\delta}) &= [\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\beta}) - \mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\sigma})\mathcal{I}^{-1}(\boldsymbol{\sigma}, \boldsymbol{\sigma})\mathcal{I}(\boldsymbol{\sigma}, \boldsymbol{\beta})]^{-1}, \\ \mathbf{W}_{12}(\boldsymbol{\delta}) &= -\mathcal{I}^{-1}(\boldsymbol{\beta}, \boldsymbol{\beta})\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\sigma}) [\mathcal{I}(\boldsymbol{\sigma}, \boldsymbol{\sigma}) - \mathcal{I}(\boldsymbol{\sigma}, \boldsymbol{\beta})\mathcal{I}^{-1}(\boldsymbol{\beta}, \boldsymbol{\beta})\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\sigma})]^{-1} = \mathbf{W}_{12}^T(\boldsymbol{\delta}), \\ \mathbf{W}_{22}(\boldsymbol{\delta}) &= [\mathcal{I}(\boldsymbol{\sigma}, \boldsymbol{\sigma}) - \mathcal{I}(\boldsymbol{\sigma}, \boldsymbol{\beta})\mathcal{I}^{-1}(\boldsymbol{\beta}, \boldsymbol{\beta})\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\sigma})]^{-1}. \end{aligned}$$

Then the generalized gradient/score vector can be expressed as

$$\tilde{S}(\boldsymbol{\delta}) = \mathbf{W}^{-1}(\boldsymbol{\delta})S(\boldsymbol{\delta}) = (\tilde{S}_1^T(\boldsymbol{\delta}), \tilde{S}_2^T(\boldsymbol{\delta}))^T, \quad (8.27)$$

where $\tilde{S}_1^T(\boldsymbol{\delta}) = \mathbf{W}_{11}(\boldsymbol{\delta})S_{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \mathbf{W}_{12}(\boldsymbol{\delta})S_{\boldsymbol{\sigma}}(\boldsymbol{\delta})$ and $\tilde{S}_2^T(\boldsymbol{\delta}) = \mathbf{W}_{21}(\boldsymbol{\delta})S_{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \mathbf{W}_{22}(\boldsymbol{\delta})S_{\boldsymbol{\sigma}}(\boldsymbol{\delta})$.

As per Remark 6.1, if the unconstrained estimate satisfies the constraints so that $\hat{\boldsymbol{\delta}} \in \mathcal{C}_1$, then the constrained estimate, $\tilde{\boldsymbol{\delta}}$, is identical to the unconstrained estimate $\hat{\boldsymbol{\delta}}$. Otherwise, we proceed with the modified GP algorithm below with an initial feasible point, $\boldsymbol{\delta}_r$, chosen from $\mathcal{C}_1$, which satisfies $\mathbf{A}\boldsymbol{\beta}_r = \mathbf{b}$. We start with the active constraint set $\mathbb{W}$ based on the constraint that holds with equality, and then we form the coefficient matrix $\bar{\mathbf{A}}$ and constraint vector $\bar{\mathbf{b}}$.

A summary of the modified GP algorithm described in Theorem 4.2 can be found in the steps

below that iterate until convergence:

Step 1) Compute $\mathbf{W}(\boldsymbol{\delta}_r)$ at the initial feasible point $\boldsymbol{\delta}_r$ of $\boldsymbol{\delta}$.

Step 2) Calculate the projection matrix $\mathbf{P}_{\mathbf{w}}(\boldsymbol{\delta}_r) = \bar{\mathbf{A}}^T [\bar{\mathbf{A}} \mathbf{W}_{11}(\boldsymbol{\delta}_r) \bar{\mathbf{A}}^T]^{-1} \bar{\mathbf{A}}$ and the distance/direction vector $\mathbf{d} = (\mathbf{d}_1^T, \mathbf{d}_2^T)^T$, where

$$\mathbf{d}_1 = [\mathbf{I} - \mathbf{W}_{11}(\boldsymbol{\delta}_r) \mathbf{P}_{\mathbf{w}}(\boldsymbol{\delta}_r)] \tilde{S}_1(\boldsymbol{\delta}_r) \quad \text{and} \quad \mathbf{d}_2 = -\mathbf{W}_{21}(\boldsymbol{\delta}_r) \mathbf{P}_{\mathbf{w}}(\boldsymbol{\delta}_r) \tilde{S}_1(\boldsymbol{\delta}_r) + \tilde{S}_2(\boldsymbol{\delta}_r).$$

Step 3) If $\mathbf{d} = \mathbf{0}$, find the Lagrange multipliers $\boldsymbol{\lambda} = [\bar{\mathbf{A}} \mathbf{W}_{11}(\boldsymbol{\delta}_r) \bar{\mathbf{A}}^T]^{-1} \bar{\mathbf{A}} \tilde{S}_1(\boldsymbol{\delta}_r)$ with component λ_i , where i is the row index of the constraint matrix $\bar{\mathbf{A}}$.

- a) If all the components of $\boldsymbol{\lambda}$ are positive, i.e. $\lambda_i \geq 0$ for $i \in \mathbb{W} \cap I_2$, associated to the active inequalities, **stop**; declare that the KT necessary conditions are satisfied at the point $\boldsymbol{\delta}_r$.
- b) If at least one component of $\boldsymbol{\lambda}$ for $i \in \mathbb{W} \cap I_2$ is negative, find the index of the smallest negative component of $\boldsymbol{\lambda}$ and remove it from the set $\mathbb{W}$. Drop a row from both $\bar{\mathbf{A}}$, and $\bar{\mathbf{b}}$, and return to **Step 2**.

Step 4) If $\mathbf{d} \neq \mathbf{0}$, search for α_1 and α_2 such that

$$\begin{aligned} \alpha_1 &= \max_{\alpha} \{ \alpha : \boldsymbol{\delta}_r + \alpha \mathbf{d} \text{ is feasible} \} \text{ and} \\ \alpha_2 &= \max_{\alpha} \{ \ell(\boldsymbol{\delta}_r + \alpha \mathbf{d}) : 0 \leq \alpha \leq \alpha_1 \}. \end{aligned}$$

Set $\tilde{\boldsymbol{\delta}}_r = \boldsymbol{\delta}_r + \alpha_2 \mathbf{d}$ and return to **Step 1**, which means that we add new constraints, if any, to $\bar{\mathbf{A}}$ and $\bar{\mathbf{b}}$ that is on the boundary to the working set $\mathbb{W}$. Then update $\boldsymbol{\delta}_r$ using $\tilde{\boldsymbol{\delta}}_r$ and return to **Step 2**.

For every step in the modified GP algorithm outlined above, the calculation of marginal likelihood, the gradient/score vector and the observed information matrix requires computation of the conditional expectation (integral) of a certain function of $\mathbf{u}$ given $\mathbf{y}$ with intermediate parameter value $\boldsymbol{\delta}_r$. These conditional expectations are approximated using numerical methods.

8.5.2 Constrained Hypothesis Tests for MGLMMs

Similar to the MGLM, we consider the constrained set $\Omega = \{\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T : \mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}\}$, which denotes the constrained parameter space, where $\mathbf{A}$ is a $r \times p$ matrix of full row rank. The constrained or restricted tests are associated with the hypotheses:

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b}, \quad H_1 : \mathbf{A}\boldsymbol{\beta} \leq \mathbf{b}, \quad H_2 : \text{no restriction on } \boldsymbol{\beta}. \quad (8.28)$$

The LRTs for the three sets of hypotheses in (8.28) are computed using the marginal log-likelihood function $\ell(\boldsymbol{\delta})$ given in Section 8.4, based on the maximum likelihood estimators, $\boldsymbol{\delta}^\circ$ under H_0 , $\tilde{\boldsymbol{\delta}}$ under H_1 , and $\hat{\boldsymbol{\delta}}$ under H_2 . Consider testing H_0 against $H_2 - H_0$. Then the unrestricted LRT statistic is given by:

$$T_{02} = 2[\ell(\hat{\boldsymbol{\delta}}) - \ell(\boldsymbol{\delta}^\circ)],$$

which asymptotically follows $\chi^2(r)$ under the null hypothesis H_0 . If T_{02} is large, then the unrestricted test rejects H_0 in favor of $H_2 - H_0$. When testing H_1 against $H_2 - H_1$, (i.e. when the parameter space is restricted by H_1), the restricted LRT statistic is given by:

$$T_{01} = 2[\ell(\tilde{\boldsymbol{\delta}}) - \ell(\boldsymbol{\delta}^\circ)].$$

When testing H_1 against $H_2 - H_1$, the restricted LRT statistic is given by:

$$T_{12} = 2[\ell(\hat{\boldsymbol{\delta}}) - \ell(\tilde{\boldsymbol{\delta}})].$$

When H_1 is true, the usefulness of the test related to the restricted LRT statistic T_{01} can be confirmed by performing the goodness of fit test, which rejects H_1 for large values of the restricted LRT statistic T_{12} . As in MGLM, the asymptotic distributions for constrained likelihood ratio tests T_{01} and T_{12} are demonstrated to be chi-bar-square, as outlined in the following theorem.

Theorem 8.1 (Asymptotic LRT distribution in MGLMM): Let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^p$ and $\mathbf{V}(\boldsymbol{\delta}_0)$ be a $p \times p$ positive definite matrix. Then under the null hypothesis H_0 , the asymptotic distributions of the LRT statistics $T_{01}(\mathbf{V}(\boldsymbol{\delta}_0), \mathcal{C})$ and $T_{12}(\mathbf{V}(\boldsymbol{\delta}_0), \mathcal{C})$ are given as follows:

$$\lim_{n \rightarrow \infty} P_{\boldsymbol{\delta}_0}\{T_{01} \geq c\} = \sum_{i=0}^q w_i(q, \mathbf{A}\mathbf{V}(\boldsymbol{\delta}_0)\mathbf{A}^T, \mathcal{C})P(\chi_i^2 \geq c), \quad (8.29)$$

$$\lim_{n \rightarrow \infty} P_{\boldsymbol{\delta}_0}\{T_{12} \geq c\} = \sum_{i=0}^q w_{q-i}(q, \mathbf{A}\mathbf{V}(\boldsymbol{\delta}_0)\mathbf{A}^T, \mathcal{C})P(\chi_i^2 \geq c), \quad (8.30)$$

for any $c \geq 0$, where q is the rank of $\mathbf{A}$, $\boldsymbol{\delta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\sigma}_0^T)^T$ is a true value of $\boldsymbol{\delta}$ under H_0 , $\mathbf{V}(\boldsymbol{\delta}_0)$ is the inverse of the Fisher information matrix, $w_i(q, \mathbf{A}\mathbf{V}(\boldsymbol{\delta}_0)\mathbf{A}^T)$ are some non-negative weights and $\sum_{i=0}^q w_i(q, \mathbf{A}\mathbf{V}(\boldsymbol{\delta}_0)\mathbf{A}^T) = 1$.

This theorem may be proved using the same approach as used in the proof of Theorem 6.1.

8.6 Constrained Statistical Inference for CCHS data

The unrestricted estimates of the random intercept multinomial logit parameters in (8.25) are presented in Table 8.1. The unrestricted estimates for $\boldsymbol{\beta}_1$ satisfy the constraints; however, the unrestricted estimates for the second set $\boldsymbol{\beta}_2$ do not satisfy the constraints under the parameter space $\mathcal{C} = \{\boldsymbol{\delta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma^2)^T : \beta_{11} \geq \beta_{21} \geq \beta_{31} \text{ and } \beta_{12} \geq \beta_{22} \geq \beta_{32}\}$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are 7×1 parameter vectors for light and moderate categories of health system use and σ^2 is the variance component.

We restrict the age groups simultaneously to have an increasing effect on the number of medical doctor consultations (i.e. light, moderate vs. heavy use of the health system). More specifically, the parameter space under H_0 is

$$\mathcal{C}_0 = \{\boldsymbol{\delta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma^2)^T : \beta_{11} = \beta_{21} = \beta_{31} \text{ and } \beta_{12} = \beta_{22} = \beta_{32}\}, \quad (8.31)$$

while under H_1 is

$$\mathcal{C}_1 = \{\boldsymbol{\delta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma^2)^T : \beta_{11} \geq \beta_{21} \geq \beta_{31} \text{ and } \beta_{12} \geq \beta_{22} \geq \beta_{32}\}. \quad (8.32)$$

To obtain the restricted estimates for the random intercept multinomial logit model in Eq. (8.25), we used the modified GP algorithm described in Section 8.5.1. This algorithm was used to find the constrained estimates for all models as shown in Table 8.2. We can rewrite the parameter space $\mathcal{C}$ in the form of pairwise contrasts $\beta_{\ell j} - \beta_{mj}$. For the light category, the constraints are:

$$\beta_{11} - \beta_{21} \geq 0, \text{ and}$$

$$\beta_{21} - \beta_{31} \geq 0.$$

For the moderate category, the constraints are:

$$\beta_{12} - \beta_{22} \geq 0, \text{ and}$$

$$\beta_{22} - \beta_{32} \geq 0.$$

Therefore, the constraint matrix can be written as:

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (8.33)$$

The hypothesis of interest are:

$$H_0 : \mathbf{A}\boldsymbol{\delta} = \mathbf{b}, \quad H_1 : \mathbf{A}\boldsymbol{\delta} \leq \mathbf{b}, \quad \text{and} \quad H_2 : \text{no restriction.}$$

Let δ^0 , $\tilde{\delta}$ and $\hat{\delta}$ be the MLE under H_0 , H_1 , and H_2 , respectively. Table 8.2 shows the MLEs under the aforementioned hypothesis. By analysing the results for the unrestricted MLEs, we can deduce that the younger the age group the higher the odds of it being in the light category versus the heavy category. Similarly, the moderate category shows a higher odd for participation of younger groups than it does for the heavy category. This is expected, as there is a proven correlation between older age and heavier use of the health care system through increased consultations with medical doctors.

When comparing the results of the unrestricted MLEs to the restricted MLEs, the effect of age on the odds of participation in each of the categories (light and moderate) versus the heavy category is confirmed and further strengthened.

Table 8.2: Unrestricted and Restricted MLE for random intercept multinomial logit of Asthma

Parameter Estimates						
# of Consultations	Covariate		Unrestricted		Restricted	
		Parameters	$\hat{\delta}$	Std. Error	δ^0	$\tilde{\delta}$
Light	Intercept	β_{01}	1.940467	0.113129	2.002481	1.939634
	age1	β_{11}	0.943946	0.131567	0.435394	0.944095
	age2	β_{21}	0.293828	0.110188	0.435394	0.356062
	age3	β_{31}	0.254779	0.104897	0.435394	0.207596
	sex	β_{41}	-0.747402	0.089372	-0.796842	-0.74522
	smoker	β_{51}	-0.389032	0.092716	-0.418969	-0.394253
	symptom	β_{61}	-0.664742	0.081414	-0.711703	-0.664217
Moderate	Intercept	β_{02}	0.706011	0.128062	0.728228	0.704317
	age1	β_{12}	0.245604	0.150985	0.015691	0.245814
	age2	β_{22}	-0.150884	0.126803	0.015691	-0.038231
	age3	β_{32}	0.044138	0.117595	0.015691	-0.038231
	sex	β_{42}	-0.328923	0.103149	-0.342095	-0.324953
	smoker	β_{52}	-0.33478	0.107929	-0.351368	-0.343624
	symptom	β_{62}	-0.162685	0.093901	-0.178156	-0.161518
Variance	regions	σ^2	0.074528	0.031609	0.067334	0.075146

The inequality constraints are noted to be significant as can be seen from the goodness-of-fit test, where the test statistic $T_{12} = 2[\ell(\hat{\boldsymbol{\delta}}) - \ell(\tilde{\boldsymbol{\delta}})] = 2[-4653.872162 + 4655.155663] = 2.567003$ with the p -value = 0.2508552. Since the alternative hypothesis does not involve equality constraints, there is no need to modify the chi-bar-square weights. For more information about computation of chi-bar-square weights, refer to Theorem 5.4.

The LRT statistic for H_0 against $H_1 - H_0$ is $T_{01} = 2[\ell(\tilde{\boldsymbol{\delta}}) - \ell(\boldsymbol{\delta}^0)] = 2[-4655.155663 + 4678.33898] = 46.366648$ with p -value < 0.000001. So we reject H_0 and conclude that age has significant and directional effects on the number of consultations with medical doctors (light, moderate versus heavy health system use), meaning that as age increases, the probability of requiring an increased number of consultations with medical doctors also increases. Moreover, the unconstrained test H_0 against $H_2 - H_1$ has LRT statistic $T_{02} = 2[\ell(\hat{\boldsymbol{\delta}}) - \ell(\boldsymbol{\delta}^0)] = 2[-4653.872162 + 4678.338988] = 48.933651$ with p -value < 0.00001, which suggests that age is significant to health system use; however, no directional effects are indicated. Therefore, the constrained LRT provides additional information which would not otherwise be detected using unconstrained hypothesis tests.

Chapter 9

Conclusion

The thesis presented here is a culmination of years of research, course work, simulations and analysis. The main theme of this thesis is to incorporate equality and inequality order restrictions on the parameters of the statistical model (multinomial logit) where we provide progressive development of related statistical theories. These theories include deriving constrained ML estimators using optimization algorithms such as the modified gradient projection algorithm and the iteratively reweighted-least squares-quadratic programming as well as the constrained LRT distribution for MGLM and MGLMM, which were shown to have an asymptotic chi-bar square distribution under the null hypothesis.

We cover a topic not yet fully researched within the field, and that has the potential to contribute significantly through proper consideration for constraints in statistical inference since this allows for increased testing power and improved accuracy in predictions. More accurate predictions can be leveraged in a number of disciplines and areas for better decision-making and policy development. To date, little research has been conducted for the advancement and development of multi-level categorical responses. This thesis addresses this problem while also expanding its usefulness with the addition of constraints, as opposed to ignoring them. For this reason, this thesis' contribution to the field through CSI in MGLM and MGLMM (multinomial logit) not only builds upon existing work GLM and GLMM (binary logit) and presents relevant findings, it also provides a foundation for future work as outlined in Section 9.3.

9.1 Main Findings

The current thesis focuses on estimating parameters while imposing constraints on them. We tested our models on simulated data sets in which the response variable is multi-categorical but was presented as count with a varying number of observations recorded on respondents. The simulation results and power comparison demonstrate efficiency gains on the models when constraints are imposed on the parameters; this confirmed our expectation of seeing improved results through the incorporation of additional constraints versus ignoring them.

These theoretical results were applied to health data from the CCHS to see if they would provide different insights from those experienced through simulations. Based on the analysis of the CCHS data, where we imposed constraints on the Age groups simultaneously to have an increasing effect on the number of medical doctor consultations, we were able to confirm our findings. It was shown that the effects of the covariates (age, sex, smoker, symptoms) are more significant in the multinomial random intercept logit model with the exception of Age3 in both light and moderate vs. heavy categories and of Age2 in the moderate category than those in the multinomial logit model with only fixed effects parameters. In other words, as expected, when Age increased, so did the number of medical doctor consultations. Our findings indicated that age was impacted in the same way across categories (increasing effect), so the younger your age, the more likely you were to belong to the light or moderate category versus the heavy.

From the simulations, we know that the unconstrained ML estimators have the following properties: they are consistent and asymptotically normally distributed, with the variance co-variance being the inverse of the Fisher information. However, for the constrained ML estimators, that property no longer holds true but the constrained estimator distribution depends on the closeness of the unconstrained estimator to the boundary of the constraint for a linear model.

9.2 Limitations

As with any research or pursuit of advancement, there are challenges and limitations. The process of researching this topic had its challenges, namely in coding and finding previous research to support the ideas presented. Some of the limitations encountered included: 1. computational complexity due to multidimensional integrals, and the complexity of the likelihood function when we introduce a random effect for each category (assuming either correlated or independent); 2. using one variance component for all categories instead of having a variance component for each of the categories; 3. sharing the same value of random effects for each category instead of adopting a more flexible approach, where categories would not share the same variance component; 4. modifying the GP algorithm for MGLM and MGLMM where the boundary for α is different; and 5. difficulty in obtaining confidence intervals in constrained environments, so hypothesis testing is preferred for inference.

However, despite the limitations, we were able to balance the need for concrete, manageable steps toward success and the time within which the work was to be completed. This balance also made way for future work that will address the limitations of the thesis. Future work in this area will further develop the ideas presented and will strengthen the conclusions made throughout this work.

9.3 Future Work

Future work in this field will delve into an aspect of ML estimators that could lead to discovering the asymptotic distribution for the constrained ML estimators. Additional inference about these estimates would flow well from this thesis into its sequel. Moreover, advances in computational capacity will simplify the study of Bayesian constrained techniques. We can consider furthering this research by finding an asymptotic approach/technique to compute the standard error for the constrained ML estimator.

Various related areas worthy of exploration can be considered, including applying constrained

inference of the MGLM and MGLMM on the missing data, testing the performance of the GP algorithm for CSI of MGLMM when we have multiple random effects (not just the random intercept), and finally, adding constraints to the variance component for the random effects and not only for the fixed effects.

List of References

- [1] A. Agresti. *Categorical Data Analysis*. Wiley, New Jersey, USA (2013).
- [2] A. Agresti. *Introduction to Categorical Data Analysis*. Wiley, New Jersey, USA (2019).
- [3] Wikipedia. “Multinomial distribution.” https://en.wikipedia.org/wiki/Multinomial_distribution (2017).
- [4] G. Rodriguez. “Multinomial response models.” <http://data.princeton.edu/wws509/notes/c6.pdf> (2007).
- [5] D. Knuth. “Knuth: Computers and typesetting.” <http://www-cs-faculty.stanford.edu/~{}uno/abcde.html>.
- [6] A. Einstein. “Zur elektrodynamik bewegter körper. (German) [On the electrodynamics of moving bodies].” *Annalen der Physik* **322**(10), 891–921 (1905).
- [7] C. R. Bilder and T. M. Loughin. *Analysis of Categorical Data with R*. Chapman and Hall/CRC Press, 6000 Broken Sound Parkway NW, Suite 300 (2015).
- [8] W. Tang, H. He, and X. M. Tu. *Applied Categorical and Count Data Analysis*. Chapman and Hall/CRC Press, 6000 Broken Sound Parkway NW, Suite 300 (2012).
- [9] D. Dawson and L. Magee. “The national hockey league entry draft, 1969-1995: An application of a weighted pool adjacent-violators algorithm.” *The American Statistician* **55**(3), 194–199 (2001).
- [10] K. Davis. *Constrained Statistical Inference in Generalized Linear, and Mixed Models with Incomplete Data*. Carleton University, Ottawa, Ontario, Canada (2011).
- [11] M. Jamshidian. “On algorithms for restricted maximum likelihood estimation.” *Computational Statistics & Data Analysis* **45**, 137–157 (2004).
- [12] R. Fletcher. *Practical Methods of Optimization*. Wiley, New York, USA (1987).
- [13] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming, 3rd edition*. Springer, New York, USA (2008).
- [14] J. Lindsey. *Applying Generalized Linear Models*. Limburgs Universitair Centrum, Diepenbeek (2007).

- [15] M. J. Silvapulle and P. K. Sen. *Constrained Statistical Inference*. Wiley (2005).
- [16] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, USA (1969).
- [17] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK (2004).
- [18] Akshita, Ramyani, Sridevi & Trishita. “Multinomial logit models, Econometrics II term paper.”
- [19] D. Bohning. “Multinomial logistics regression algorithm.” *Annals of the Institute of Statistical Mathematics* **40**(1), 197–200 (1992).
- [20] Y. Li, W. Gao, and N.-Z. Shi. “A note on multinomial maximum likelihood estimation under ordered restrictions and the EM algorithm.” *Metrika* **66**(1), 105–114 (2007).
- [21] A. Hasan, Z. Wang, and A. S. Mahani. “Fast estimation of multinomial logit models: R package mnlogit.” *Journal of Statistical Software* **75**(3), 1–24 (2016).
- [22] Y. Croissant. “Estimation of multinomial logit models in R : The mlogit packages.”
- [23] D. Hosmer Jr., S. Lemeshow, and R. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Inc., Hoboken, New Jersey (2013).
- [24] L. A. Thompson. *R (and S-Plus) Manual to Accompany Agresti’s Categorical Data Analysis (2002)*. publisher unknown (2009).
- [25] J. K. Dow and J. W. Endersby. “Multinomial probit and multinomial logit: a comparison of choice models for voting research.” *Electoral Studies* **23**(1), 107–222 (2004).
- [26] S. A. Czepiel. “Maximum likelihood estimation of logistic regression models: Theory and implementation.” <https://czep.net/stat> (2019).
- [27] K. A. Davis, C. G. Park, and S. K. Sinha. “Testing for generalized linear mixed models with cluster correlated data under linear inequality constraints.” *The Canadian Journal of Statistics* **40**(2), 243–258 (2012).
- [28] J. J. Faraway. *Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press (2016).
- [29] Statistics Canada. “Health regions and peer groups.” <https://www150.statcan.gc.ca/n1/pub/82-402-x/2015001/regions/hrpg-eng.htm>.

- [30] Statistics Canada. “Canadian community health survey - annual component (C-CHS) 2018.” <http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=795204>.
- [31] Statistics Canada. “Canadian community health survey - annual component (C-CHS) 2012.” <http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=135927>.
- [32] UCLA. “Multinomial logistic regression — SPSS annotated output.” <https://stats.idre.ucla.edu/spss/output/multinomial-logistic-regression/> (2019).
- [33] G. Huthcheson. “Modelling ordered and unordered categorical data using logit models.” <https://pdfs.semanticscholar.org/463e/b9ae434762bd58ec68f44efb93e80d9b4797.pdf> (2019).
- [34] T. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley (2003).
- [35] T. Anderson. *The Statistical Analysis of Time Series*. Wiley (1994).
- [36] T. W. Anderson. “The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities.” *Proceedings of the American Mathematical Society* **6**(2), 170–176 (1955).
- [37] D. W. K. Andrews. “Hypothesis testing with a restricted parameter space.” *Journal of Econometrics* **84**, 155–199 (1998).
- [38] C. E. McCullough and S. R. Searle. *Generalized, Linear, and Mixed Models*. Wiley (2001).
- [39] W. W. Stroup. *Generalized, Linear, and Mixed Models. Modern Concepts, Methods and Applications*. Chapman & Hall, CRC Press (2013).
- [40] J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer (2007).
- [41] E. Demidenko. *Mixed Models: Theory and Applications with R, Second edition*. John Wiley & Sons (2013).
- [42] P. McCullagh and J. Nelder. *Generalized Linear Models, Second edition*. Chapman and Hall (1989).
- [43] J. M. Hilbe. *Logistic Regression Models*. Chapman & Hall, CRC Press (2009).
- [44] R. C. Rao. *Linear Statistical Inference and its Applications, Second edition*. John Wiley & Sons (2002).

- [45] G. Tutz. *Regression for Categorical Data*. Cambridge University Press (2012).
- [46] D. G. Luenberger. *Linear and NonLinear Programming. Second edition*. Addison-Wesley Publishing Company (1984).
- [47] A. Rothwell. *Optimization Methods in Structural Design*. Springer (2017).
- [48] K. K. Choi and N. H. Kim. *Structural Sensitivity Analysis and Optimization, Linear Systems*. Springer (2005).
- [49] K. K. Choi and N. H. Kim. *Structural Sensitivity Analysis and Optimization 2, Nonlinear Systems and Applications*. Springer (2005).
- [50] N. H. Kim, D. An, and J.-H. Choi. *Prognostics and Health Management of Engineering Systems, An Introduction*. Springer (2017).
- [51] N. H. Kim. *Introduction to Nonlinear Finite Element Analysis*. Springer (2015).
- [52] G. H. Golub and C. F. Van Loan. *Matrix Computations. Third edition*. The Johns Hopkins University Press (1996).
- [53] M. Kry and A. Royle. *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press (2015).
- [54] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity. The Lasso and Generalizations*. Chapman & Hall, CRC Press (2016).
- [55] J. H. McDonald. *Handbook of Biological Statistics. Third edition*. Sparky House Publishing. (2014).
- [56] N. G. Becker. *Modeling to Inform Infectious Disease Control*. Chapman & Hall, CRC Press (2015).
- [57] W. N. Venables and B. Ripley. *Modern Applied Statistics with S. Fourth edition*. Springer (2008).
- [58] J. Hwang and S. Peddada. “Confidence interval estimation subject to order restrictions.” *The Annals of Statistics* **22**, 67–93 (1994).
- [59] D. Dunson and B. Neelon. “Bayesian inference on orderconstrained parameters in generalized linear models.” *Journal of the International Biometric Society* **59**, 286–295 (2003).
- [60] J. Clavin and R. Dykstra. “Reaml estimation of covariance matrices with restricted parameter spaces.” *Journal of the American Statistical Association* **90**, 321–329 (1995).

- [61] L. Fahrmeir and H. Kaufmann. “Consistency and asymptotic of the maximum likelihood estimator in generalized linear models.” *The Annals of Statistics* **13**, 342–368 (1985).
- [62] K. A. Davis. “Constrained statistical inference: A hybrid of statistical theory, projective geometry and applied optimization techniques.” *Progress in Applied Mathematics* **4**, 167–181 (2012).
- [63] Wikipedia. “Constance van eeden.” https://en.wikipedia.org/wiki/Constance_van_Eeden (2019).
- [64] Wikipedia. “Isotonic regression.” https://en.wikipedia.org/wiki/Isotonic_regression (2019).
- [65] H. E. Barmi and R. L. Dykstra. “Maximum likelihood estimates via duality for log-convex models when cell probabilities are subject to convex constraints.” *The Annals of Statistics* **26**, 1878–1893. https://projecteuclid.org/download/pdf_1/euclid.aos/1024691361 (1998).
- [66] X. Lin. “Variance component testing in generalised linear models with random effects.” *Biometrika* **84**, 309326. <https://academic.oup.com/biomet/article-abstract/84/2/309/233889> (1997).
- [67] D. B. Hall and J. T. Prstgaard. “Orderrestricted score tests for homogeneity in generalised linear and nonlinear mixed models.” *Biometrika* **88**, 739751. <https://academic.oup.com/biomet/article-abstract/88/3/739/340111?redirectedFrom=fulltext> (2001).
- [68] A. Kudo. “A multivariate analogue of the one-sided test.” *Biometrika* **50**, 403418. <https://doi.org/10.1093/biomet/50.3-4.403> (1963).
- [69] M. J. Silvapulle. “On tests against one-sided hypotheses in some generalized linear models.” *Biometrics* **50**, 853–858. <https://www.jstor.org/stable/2532799> (1994).
- [70] Y. Lee and J. A. Nelder. “Hierarchical generalized linear models.” *Journal of the Royal Statistical Society. Series B* **54**, 619–678 (1996).
- [71] A. Agresti, J. G. Booth, J. P. Hobert, and B. Caffo. “Random effects modeling of categorical response data.” *Sociological Methodology* **30**, 27–80 (2008).
- [72] G. Tutz and A. Groll. “Binary and ordinal random effects models including variable selection.” *Institut for Statistik* **97**, 27–80 (2010).
- [73] G. Tutz and W. Hennevogl. “Random effects in ordinal regression models.” *Computational Statistics and Data Analysis* **22**, 537–557 (1996).

- [74] D. Hedeker. “A mixed effects multinomial logistic regression model.” *Statistics in Medicine* **22**, 1433–1446 (2003).
- [75] G. Papageorgiou and J. Hinde. “Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities.” *Statistics and Computing* **22**, 79–92 (2012).
- [76] A. A. Jonathan Hartzel and B. Caffo. “Multinomial logit random effects models.” *Statistical Modelling* **1**, 81–102 (2001).
- [77] Z. Chen and L. Kuo. “A note on the estimation of the multinomial logit model with random effects.” *The American Statistician* **2**, 89–95 (2001).
- [78] I. Das and S. Mukhopadhyay. “On generalized multinomial models and joint percentile estimation.” *Journal of Statistical Planning and Inference* **145**, 190–203 (2014).
- [79] G. Glasgow. “Mixed logit models in political science.” <http://www.polsci.ucsb.edu/faculty/glasgow> (2001).
- [80] B. A. Coull and A. Agresti. “Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution.” *Biometrics* **56**, 73–80 (2000).
- [81] R. Schall. “Estimation in generalized linear models with random effects.” *Biometrika* **78**, 719–727 (1991).
- [82] N. Malchow-Mller and M. Svarer. “Estimation of the multinomial logit model with random effects.” *Applied Economics Letters* **10**, 389–392 (2003).
- [83] C. R. Bhat. “Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model.” *Pergamon* **35**, 677–693 (2001).
- [84] O. Lukociene and J. K. Vermunt. “Logistic regression analysis with multidimensional random effects: A comparison of three approaches.” PhD Dissertation (2009).
- [85] J. Hinde. “Compound poisson regression models.” *Lecture Notes in Statistics* **14**, 109–121 (1982).
- [86] D. Anderson and M. Aitkin. “Variance component models with binary response: Interviewer variability.” *Journal of the Royal Statistical Society. Series B* **47**, 203–210 (1985).
- [87] S. L. Zeger and M. R. Karim. “Generalized linear models with random effects; a gibbs sampling approach.” *Journal of the American Statistical Association* **86**, 79–86 (1991).
- [88] R. McCulloch and P. E. Rossi. “An exact likelihood analysis of the multinomial probit model.” *Journal of Econometrics* **64**, 207–240 (1994).

- [89] R. McCulloch. “Maximum likelihood algorithms for generalized linear mixed models.” *Journal of the American Statistical Association* **92**, 162–170 (1997).
- [90] J. G. Booth and J. P. Hobert. “Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm.” *Journal of the Royal Statistical Society. Statistical Methodology, Series B* **61**, 265–285 (1999).
- [91] N. E. Breslow and D. G. Clayton. “Approximate inference in generalized linear mixed models.” *Journal of the American Statistical Association* **88**, 9–25 (1993).
- [92] R. Wolfinger and M. O’Connell. “Generalized linear mixed models a pseudo-likelihood approach.” *Journal of Statistical Computation and Simulation* **48**, 233–243 (1993).
- [93] R. Stiratelli, N. Laird, and J. Ware. “Random-effects models for serial observations with binary response.” *Biometrics* **40**, 961–971 (1984).
- [94] A. Gilmour, R. Anderson, and A. Rae. “The analysis of binomial data by generalized linear mixed model.” *Biometrika* **72**, 539–599 (1985).
- [95] N. Breslow. “Extra-poisson variation in log-linear models.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **33**, 38–44 (1984).
- [96] D. Harville and R. Mee. “A mixed-model procedure for analyzing ordered categorical data.” *Biometrics* **40**, 393–408 (1984).
- [97] J. J. “On the statistical analysis of ordinal data when extravariation is present.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **39**, 74–85 (1990).
- [98] F. Ezzet and J. Whitehead. “A random effects model for ordinal responses from a crossover trial.” *Statistics in medicine* **10**, 901–906 (1991).
- [99] D. Hedeker and R. Gibbons. “A random-effects ordinal regression model for multilevel analysis.” *Biometrics* **50**, 933–944 (1994).
- [100] J. B. Lang. “On the comparison of multinomial and poisson log-linear models.” *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. **58**, 253–266 (1996).
- [101] P. Mizdrak. *Clustering Profiles in Generalized Linear Mixed Models Settings Using Bayesian Nonparametric Statistics*. Carleton University, Ottawa, Ontario, Canada (2018).

Appendix A

Optimization Algorithms

A.1 Introduction

Let $\hat{\beta}$ be the MLE that can be computed by numerical procedures (computing iteratively until convergence) since there is no closed form solution for optimal β . There are three types of iterative algorithms: the **Expectation Maximization (EM)**, **Newton-Raphson (NR)** and **Fisher Scoring (FS)**. These algorithms differ by a negative Hessian matrix $\mathbf{H}$, which is the second derivative of the log-likelihood function with respect to β . NR and FS are the preferred algorithms for Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) estimation because they have a quadratic convergence and produce an asymptotic covariance matrix $\mathbf{H}^{-1}$ of the estimated parameter. Statistically, we prefer to use the FS because it uses the expected negative Hessian matrix to estimate the covariance parameter while the NR only uses the observed Hessian matrix. Also quasi-likelihood methods can be used which specify only the mean and variance relationship, rather than the full likelihood.

A.2 The Newton-Raphson Method

To find the minimum or maximum of a function, we often use the Newton-Raphson (NR) algorithm. We are focusing on the multinomial logit model, where we find the maximum of $\ell(\beta)$ which occurs when the gradient/score of $\ell(\beta)$ is equal to the zero matrix, meaning when $\frac{\partial}{\partial \beta} \ell(\beta) = \mathbf{0}$. This means that the maximum is attained when the gradient points are equal to zero. The NR method brings us to the nearest point to the maximum; this method is called local optimization. To obtain good results using local optimization, we must choose a good

initial starting value β_0 . As we already know, the NR method can be explained as a loop in which two steps are used: 1) we iterate and find new values for the coefficients, and 2) we test for convergence. In reference to the multinomial logit model, we use the NR algorithm in this fashion:

- (1) For $t = 0, \dots$ update the current position using

$$vec(\beta_{t+1}) = vec(\beta_t) + \mathbf{H}_t^{-1}vec(s_t), \quad (\text{A.1})$$

where s_t represents the gradient/score points at position t and their position in relation to the maximum. $\mathbf{H}_t^{-1}$ is evaluated at β_t and gives information about the curvature of the log-likelihood, thereby identifying the rate at which one could reach the maximum.

- (2) The NR algorithm is repeated until s_t is as close as possible to $\mathbf{0}$.

However, it is important to note two scenarios in which convergence may not be possible; these scenarios should be accounted for and workarounds should be implemented. The first scenario is one in which a model is not well-defined, causing a parameter estimate to tend toward infinity; this depends on the initial starting value chosen. The second scenario is one in which an estimate overshoots the root, causing a repeated cycle of iterations that will never converge. To avoid the first scenario, the initial value (the least square estimate $\hat{\beta}_0$) was obtained by regressing the $\ln\left(\frac{\mathbf{Y}}{\mathbf{n}} + 0.1\right)$ on the covariate matrix $\mathbf{X}$. To avoid the second scenario, we introduce sub-iterations, whereby we implement “step-halving”.

Appendix B

Exponential Family

A class of distributions, ranging between both continuous and discrete random variables, with the following form are part of the one-parameter exponential family (EF):

$$f(y_i; \eta) = h(y_i)s(\eta_i)\exp\{T(y_i)u(\eta_i)\}, \quad (\text{B.1})$$

where Y_i is an independent response variable and η_i is a location parameter, which indicates the location of the distribution in the range of possible response values, and $i = 1, \dots, n$. To simplify the above, we obtain the family distribution, the parameter and the canonical form by performing a one-to-one transformation $x = t(y)$ and $\theta = u(\eta)$.

This can be rewritten in the following canonical form, where $a(\theta_i)$ is a normalizing constant distribution:

$$f(x_i; \theta_i) = \exp\{x_i\theta_i - a(\theta_i) + c(x_i)\}. \quad (\text{B.2})$$

If we then add a scale parameter, ϕ , to the above, we can generalize the exponential dispersion family to give the following:

$$f(x_i; \theta_i, \phi) = \exp\left\{\frac{x_i\theta_i - a(\theta_i)}{b_i(\phi)} + c(x_i, \phi)\right\}, \quad (\text{B.3})$$

where θ_i remains the canonical form of η_i , a function of the mean μ_i .

The mean and variance of the exponential and exponential dispersion families hold a special relationship. The likelihood function $L(\theta, \phi; x) = \prod_{i=1}^n f(x_i; \theta_i, \phi)$ is one method in which we can obtain the variance and the mean, for which the first derivative of the log likelihood

$\ell(\theta, \phi; x) = \log \left(L(\theta, \phi; x) \right)$ is obtained by:

$$U = \frac{\partial \ell}{\partial \theta}. \quad (\text{B.4})$$

If we set the equation (B.4) to zero, then the MLE is derived.

From the standard inference theory, we can show that:

$$\text{E}(U) = 0 \quad \text{and} \quad \text{V}(U) = \text{E}(U^2) = \text{E} \left(-\frac{\partial U}{\partial \theta} \right). \quad (\text{B.5})$$

The log likelihood for a particular observation $\ell(\theta_i, \phi; x_i) = \frac{x_i \theta_i - a(\theta_i)}{b_i(\phi)} + c(x_i, \phi)$, then for each θ_i ,

$$U_i = \frac{X_i - \frac{\partial a(\theta_i)}{\partial \theta_i}}{b_i(\phi)}. \quad (\text{B.6})$$

From equation (B.5),

$$\text{E}(U_i) = \text{E} \left(\frac{X_i - \frac{\partial a(\theta_i)}{\partial \theta_i}}{b_i(\phi)} \right) = 0, \quad (\text{B.7})$$

so that

$$\text{E}(X_i) = \frac{\partial a(\theta_i)}{\partial \theta_i} = \mu_i. \quad (\text{B.8})$$

From equation (B.5), $U'_i = -\frac{\frac{\partial^2 a(\theta_i)}{\partial \theta_i^2}}{b_i(\phi)}$. Then, the variance of U_i is obtained using the formula (B.6):

$$\text{V}(U_i) = \frac{\text{V}(X_i)}{b_i^2(\phi)} = \frac{\partial^2 a(\theta_i)}{\partial \theta_i^2} b_i(\phi). \quad (\text{B.9})$$

Then, we rearrange the above equation, which yields $\text{V}(X_i) = \frac{\partial^2 a(\theta_i)}{\partial \theta_i^2} b_i(\phi)$. This can be further simplified by taking $b_i(\phi) = \frac{\phi}{w_i}$, where w_i represents prior weights. If we then let the variance function (a function of μ_i or θ_i only) $\frac{\partial^2 a(\theta_i)}{\partial \theta_i^2} = \tau_i^2$, we obtain the product of the dispersion parameter and a function of the mean. The derivation method described in this appendix for

the EF was obtained largely from J.K. Lindsey [14].

$$V(X_i) = b_i(\phi)\tau_i^2 = \frac{\phi\tau_i^2}{w_i} \quad (\text{B.10})$$

The EF members share the same properties.

- The product of the pdf for two random variables (X, Y) , or more, will belong to an EF if the pdf for each of the random variables belongs to an EF.
- Bayesian estimation is easy to calculate because every EF distribution has a conjugate prior.
- For the modelling purpose, if Y is from an EF, then $V(Y) = V(\mu)\phi$, where V is a known function of $\mu = E(Y)$, and ϕ is a scale parameter.

Table B.1 lists distributions that are members of the EF:

Table B.1: Exponential Distribution

Exponential Distribution	Domain
Bernoulli	binary{0,1}
Beta	(0,1)
Binomial	counts of success or failure
Dirichlet	(Simplex)
Exponential	$\mathbb{R}^+$
Gamma	$\mathbb{R}^+$
Gaussian	$\mathbb{R}^p$
Laplace	$\mathbb{R}^+$
Multinomial	<i>categorical</i>
Poisson	$\mathbb{N}^+$
Von mises	sphere
Weibull	$\mathbb{R}^+$
Weishart	symmetric positive definite matrices

The lognormal and Pareto distributions are not in the exponential family.

Appendix C

Linear Spaces

This appendix discusses the basic properties of vector space and normed linear spaces. A normed linear space is a vector space having a measure of distance or length defined on it. With the introduction of a norm, it becomes possible to define analytical or topological properties such as convergence and open and closed sets. This appendix is directly referenced from [16].

C.1 Vector Spaces

Definition C.1 (Vector Space): A vector space X is a set of elements called vectors with two operations, addition and scalar multiplication, satisfying:

$$x, y \in X \Rightarrow x + y \in X \text{ and } \alpha x \in X \quad \forall \text{ scalar } \alpha.$$

These operations are assumed to satisfy the following axioms:

- (1) $x + y = y + x$.
- (2) $(x + y) + z = x + (y + z)$.
- (3) There is a null vector 0 in X such that $x + 0 = x$ for all x in X .
- (4) $\alpha(x + y) = \alpha x + \alpha y$
- (5) $(\alpha + \beta)x = \alpha x + \beta x$
- (6) $(\alpha\beta)x = \alpha(\beta x)$
- (7) $0x = 0, 1x = x$

C.1.1 Subspaces, Linear Combinations, and Linear Varieties

Definition C.2 (Subspaces): A nonempty subset M of a vector space X is called a subspace of X if $X, Y \in M$, then the vector $\alpha x + \beta y \in M \quad \forall x, y \in \mathbb{R}$.

Proposition C.1: Let M and N be subspaces of the same dimension of a vector space X . Then:

- (1) sum, $M + N$, is a subspace of X .
- (2) intersection, $M \cap N$, is a subspace of X .

Definition C.3: Suppose S is a subset of a vector space X . Then the span of S , denoted by $[S]$, called the *subspace generated* by S , is the set that consists of all possible linear combinations of vectors in S .

Definition C.4 (Affine Subspace): The translation of a subspace is said to be a *linear variety*. If M is subspace of X i.e. $M \subset X$ then a linear variety $V = M + x_0 = \{m + x_0 : m \in M\}$. A linear variety is also known as flat, affine set, affine subspace, and linear manifold.

Definition C.5: $A \subset X$ is an affine set if $x, y \in A \Rightarrow \lambda x + (1 - \lambda)y \in A \quad \forall \lambda \in \mathbb{R}$.

Definition C.6: Affine sets A and B are parallel if $A = B + x_0$ for some $x_0 \in X$.

$$x, y \in A \Rightarrow \lambda x + (1 - \lambda)y \in A \quad \forall \lambda \in \mathbb{R}$$

Proposition C.2: The following propositions are true:

- (1) The subspaces are the affine sets which contain origin.
- (2) Each non-empty affine set is parallel to a unique subspace, which conclude that the affine set and the linear varieties are the same.
- (3) The intersection of linear varieties is a linear variety.

Definition C.7: Let S be a nonempty set in X . Then the linear variety generated by S is the smallest linear variety containing S (i.e. the intersection of all linear variety containing S).

Figure C.1: Cone

C.1.2 Convexity and Cones

Definition C.8 (Convex): A set K in a linear vector space is said to be *convex* if, given $x, y \in K$ then all point of the form $\alpha x + (1 - \alpha)y \in K \quad \forall 0 \leq \alpha \leq 1$.

Proposition C.3: Let K_1 and K_2 be convex sets in a vector space. Then:

- (1) Subspace is convex.
- (2) $\alpha K_1 = \{x : x = \alpha k, k \in K_1\}$ is convex for any scalar α .
- (3) The sum $K_1 + K_2$ and the intersection $K_1 \cap K_2$ are convex.

Proposition C.4: Let $\mathcal{C}$ be an arbitrary collection of convex sets. Then $\bigcap_{K \in \mathcal{C}} K$ is convex.

Definition C.9: Let $S \subset X$. Then *convex cover* or *convex hull*, denoted $Co(S)$, is the smallest convex set containing S . In other words, $Co(S)$ is the intersection of all convex sets containing S .

Definition C.10 (Cone): A set K in a linear vector space is said to be a cone with vertex at the origin if $x \in K$ implies that $\alpha x \in K$ for all $\alpha \geq 0$.

Definition C.11 (Hyper-planes and half-spaces): Let $\mathbf{x} \in \mathbb{R}^p$, then:

- (1) **Hyper-plane:** set of the form $\{\mathbf{x} : \mathbf{a}^T \mathbf{x} = 0\}$, where $\mathbf{a} \neq \mathbf{0}$.
- (2) **Half-plane:** set of the form $\{\mathbf{x} : \mathbf{a}^T \mathbf{x} \leq 0\}$, where $\mathbf{a} \neq \mathbf{0}$.

$\mathbf{a}$ is the normal vector, hyper-planes are affine and convex; half-spaces are convex.

Proposition C.5: *The following propositions are true:*

- (1) K is a convex cone iff $x, y \in K \Rightarrow \alpha x + \beta y \in K \quad \forall \alpha, \beta \geq 0$.
- (2) $K = \{x \in \mathbb{R}^p : x_1 \leq x_2 \leq \cdots \leq x_p\}$ is a convex cone.
- (3) $K = \{x \in \mathbb{R}^p : \sum_{j=1}^p a_j x_j \leq 0\}$ is a convex cone.
- (4) $K = \{x \in \mathbb{R}^p : \sum_{j=1}^p a_j x_j \leq b\}$ is a translated cone.
- (5) A convex cone K is a **proper cone** if:
 - K is closed (contains its boundary),
 - K is solid (has nonempty interior), or
 - K is pointed (contains no line).

C.1.3 Linear Independence and Dimension

Definition C.12 (Linear dependent): A vector x is said to be linearly dependent on a set S of vectors if $x \in [S]$, the subspace generated by S . Equivalently, x can be expressed as a linear combination of vectors from S .

Proposition C.6: A set of vectors $x_1, \dots, x_n$ are said to be a linearly independent iff $\sum_{i=1}^n c_i x_i = 0$ implies that $c_1 = \cdots = c_n = 0$

Definition C.13 (Basis): A finite set S of linearly independent vectors is said to be a **basis** for the space X if S generates X . A vector space having a finite basis is said to be **finite dimensional** (i.e. Dimension of $X = \#(S)$). All other vector spaces are said to be **infinite dimensional**.

C.2 Normed Linear Space

Definition C.14 (Normed X): A vector space X is called **normed linear vector space** if there is a defined real-valued function $\|\cdot\|$, called norm of x , defined on X . This function maps each vector x in X into a real number. The norm satisfies the following axioms:

- (1) $\|x\| \geq 0 \quad \forall x \in X$ and $\|x\| = 0$ iff $x = 0$.
- (2) Triangle inequality if $x, y \in X$, then $\|x + y\| \leq \|x\| + \|y\|$.
- (3) For every $\alpha \in \mathbb{R}$ and $x \in X$, $\|\alpha x\| = |\alpha| \|x\|$.

Note: Using the triangle inequality we can show that $\|x\| - \|y\| \leq \|x - y\|$.

Example C.1: If $X = C[a, b] = \{x : x(t) \text{ can be continuous on } [a, b]\}$ is normed linear vector space with the norm $\|x\| = \max_{a \leq t \leq b} |x(t)|$.

Note: $X = C[a, b]$ can be another normed vector space with the norm $\|x\| = \int_a^b |x(t)| dt$.

Example C.2: If $X = D[a, b] = \{x : x(t) \text{ can be continuous and have continuous derivatives on } [a, b]\}$ is normed linear vector space. The norm on space $D[a, b]$ is defined as

$$\|x\| = \max_{a \leq t \leq b} |x(t)| + \max_{a \leq t \leq b} |x'(t)|.$$

Example C.3 (Euclidean Space): Euclidean space, denoted E^n , is composed of n -tuples $x = (x_1, \dots, x_n)$ with the norm defined as $\|x\| = \left(\sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}}$.

Definition C.15 (ℓ_p Space): Let p be a real number $1 \leq p \leq \infty$. The ℓ_p space is the set composed of all sequence of scalars $\{\xi_1, \xi_2, \dots\}$ for which $\sum_{i=1}^{\infty} |\xi_i|^p < \infty$. The norm of vector $x = \{\xi_1, \xi_2, \dots\}$ in ℓ_p is defined as

$$\|x\|_p = \left(\sum_{i=1}^n |\xi_i|^p \right)^{\frac{1}{p}}.$$

Note: The space ℓ_{∞} consists of bounded sequences. The norm of vector $x = \{\xi_1, \xi_2, \dots\}$ in ℓ_{∞} is defined as

$$\|x\|_{\infty} = \sup_i |\xi_i|.$$

Definition C.16: The space $L_p[a, b]$ consists of those real-valued measurable functions $x(t)$ on the interval $[a, b]$ for which $|x(t)|^p$ is Lebesgue integrable. i.e.

$$L_p[a, b] = \{x : x(t) \text{ is a Lebesgue Integrable function on } [a, b]\}.$$

The norm of this space is defined as

$$\|x\|_p = \left(\int_a^b |x(t)|^p \right)^{\frac{1}{p}}.$$

Note: On this space $\|x\|_p = 0$ does not imply $x = 0$ (i.e. $x(t)$ may be nonzero on a set of measure zero).

C.2.1 Open and Closed Sets

Let $S_\epsilon(x) = \{y : \|y - x\| < \epsilon\}$ be open sphere (ball) centered at x with radius $\epsilon > 0$

Definition C.17: Let P be a subset of a normed space X . The point $p \in P$ is an interior point of P if *there is* an $\epsilon > 0$ such that all vectors x satisfying $\|x - p\| < \epsilon$ are also members of P , i.e. $S_\epsilon(x) \subset P$.

Note: $\mathring{P}$ is called the interior of P , which is the set of all interior points of P .

Definition C.18: A point $x \in X$ is a closure point of a set P if $\exists p \in P$ such that $\|x - p\| < \epsilon \quad \forall \epsilon > 0$. This means that, a point x is a closure point of P if every sphere centered at x contains a point of P .

Note: $\bar{P}$ is called the closure of P , which is the set of all closure points of P . It is clear that $P \subset \bar{P}$.

Note C.1: *The following is true:*

(1) A set P is said to be open if $P = \mathring{P}$.

(2) A set P is said to be closed if $P = \bar{P}$.

(3) The complement of an open set is closed and the complement of a closed set is open.

(4) If K is a convex set in normed space, then $\bar{K}$ and $\overset{\circ}{K}$ are convex.

C.2.2 Banach Spaces

Definition C.19: A sequence $\{x_n\}$ in a normed vector space is said to be a Cauchy sequence if $\|x_n - x_m\| \rightarrow 0$ as $n, m \rightarrow \infty$ i.e., given $\epsilon > 0$, there is an integer N such that $\|x_n - x_m\| < \epsilon$ for all $n, m > N$.

Definition C.20 (Banach Space): A normed linear vector space X is complete if every Cauchy sequence from X has a limit in X . A complete normed linear vector space is called a *Banach space*.

Example C.4: The space R^n with norm $\|x\|^2 = \sum_{i=1}^n x_i^2 w_i$ where $w_i > 0 \quad \forall i$ is complete.

Example C.5: The space $C[a, b]$ with norm $\|x\| = \sup_{a \leq t \leq b} |x(t)|$ is complete.

Note: X is a space of continuous function on $[0, 1]$ with norm $\|x\| = \int_0^1 |x(t)| dt$ is not complete.

Let

$$x_n(t) = \begin{cases} 0 & \text{if } 0 \leq t < \frac{1}{2} - \frac{1}{n} \\ nt - \frac{n}{2} + 1 & \text{if } \frac{1}{2} - \frac{1}{n} \leq t < \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq t \leq 1 \end{cases},$$

$\|x_n - x_m\| = \frac{1}{2} \left| \frac{1}{n} - \frac{1}{m} \right| \rightarrow 0$ as $n, m \rightarrow \infty$ but

$$x_n(t) \rightarrow x(t) = \begin{cases} 0 & \text{if } 0 \leq t < \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq t \leq 1 \end{cases}.$$

Example C.6: The space $\ell_p = \{x : x = \{\xi_i\}_{i=1}^\infty \text{ with } \sum_{i=1}^\infty |\xi_i|^p < \infty\}$ with norm $\|x\|_p = \left(\sum_{i=1}^n |\xi_i|^p \right)^{\frac{1}{p}}$ is a Banach space, where $1 \leq p \leq \infty$.

Example C.7: The space $L_p[0, 1]$ of Lebesgue integrable functions on $[0, 1]$ with norm $\|x\|_p =$

$\left(\int_a^b |x(t)|^p\right)^{\frac{1}{p}}$ is a Banach space.

C.3 Hilbert Space

Hilbert Space are Banach spaces with a norm that is derived from inner product, so they have extra feature in comparison with arbitrary Banach spaces, which makes them still more special.

Definition C.21: Let X be a (linear) vector space. An inner product on $X \times X$ is a real-valued function

$$\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{R} \text{ such that}$$

- (i) $\langle x, y \rangle = \langle y, x \rangle$,
- (ii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$,
- (iii) $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle \quad \forall \lambda \in \mathbb{R}$,
- (iv) $\langle x, x \rangle \geq 0, \langle x, x \rangle = 0 \iff x = 0$.

A vector space together with $\langle \cdot, \cdot \rangle$ is called a pre-Hilbert space.

Lemma C.1 (Cauchy Schwartz Inequality): $|\langle x, y \rangle| \leq \|x\| \|y\| \quad \forall x, y \in$
inner product space where $\|x\| = \sqrt{\langle x, x \rangle}$.

Proof. Suppose $y = 0 = 0.x \quad \forall x$ trivial

$$0 \leq \langle x - \lambda y, x - \lambda y \rangle = \langle x, x \rangle - 2\lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle \quad \forall \lambda,$$

therefore $\langle x, y \rangle^2 - \langle x, x \rangle \langle y, y \rangle \leq 0 \Rightarrow |\langle x, y \rangle| \leq \|x\| \|y\|$ □

Example C.8: The following are examples of pre-Hilbert space:

- (1) $X = \mathbb{R}^n$ with $\langle x, y \rangle = \sum_{i=1}^n w_i x_i y_i$ where $w_i > 0 \quad \forall i$ and $\|x\| = \sqrt{\sum_{i=1}^n w_i x_i^2}$

(2) ℓ_2 -space = $\{x : x = \{\xi_i\}_{i=1}^\infty \text{ with } \sum_{i=1}^n |\xi_i|^2 < \infty\}$ with $\langle x, y \rangle = \sum_{i=1}^n \xi_i \eta_i$

(3) $L_2[a, b] = \{x : x(\cdot) \text{ is Lebesgue integrable with } \int_a^b |x(t)|^2 dt < \infty\}$ with $\langle x, y \rangle = \int_a^b x(t)y(t)dt$

Lemma C.2 (Parallelogram Law): $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \forall x, y \text{ in a pre-Hilbert space.}$

Definition C.22: If $\langle x, y \rangle = 0$, x is said to be orthogonal to y denoted as $(x \perp y)$.

Definition C.23: A complete pre-Hilbert space is a Hilbert space.

Lemma C.3: Let X be Hilbert space. Suppose $x_n, y_n \in X \quad \forall n$. If $x_n \rightarrow x$ and $y_n \rightarrow y$ then $\langle x_n, y_n \rangle \rightarrow \langle x, y \rangle$, the continuity of inner product.

Proof. Let $x_n = x_n - x + x \Rightarrow \|x_n\| = \|x_n - x\| + \|x\|$, and since $x_n \rightarrow x$, $\exists k$ such that $\|x_n - x\| < k$. Therefore $\|x_n\| < M$ for some $M < \infty$.

$$\begin{aligned}
 |\langle x_n, y_n \rangle - \langle x, y \rangle| &= |\langle x_n, y_n \rangle - \langle x_n, y \rangle + \langle x_n, y \rangle - \langle x, y \rangle| \\
 &= |\langle x_n, y_n - y \rangle + \langle x_n - x, y \rangle| \\
 &\leq |\langle x_n, y_n - y \rangle| + |\langle x_n - x, y \rangle| \\
 &\leq \|x_n\| \|y_n - y\| + \|x_n - x\| \|y\| \rightarrow 0
 \end{aligned}$$

□

Appendix D

Big O and Small o

Big O and small o notation is useful for describing limiting behavior of sequences. The below definitions for O_p and o_p are from Jiang, Jiming 2007 [40] and Alan Agresti [1].

Definition D.1 (Big O_p): A sequence of random vectors (including random variables), $\boldsymbol{\xi}_n$, is said to be bounded in probability, denoted by $O_P(1)$, if for any $\epsilon > 0$, there is $\delta > 0$ such that $pr(|\boldsymbol{\xi}_n| > \delta) < \epsilon$, $n = 1, 2, \dots$. If a_n is a sequence of positive numbers, the notation $\boldsymbol{\xi}_n = O_p(a_n)$ means that $\frac{\boldsymbol{\xi}_n}{a_n} = O_p(1)$. For instance, $\frac{3}{n} + \frac{8}{n^2}$ is $O(n^{-1})$ as $n \rightarrow \infty$; dividing it by n^{-1} gives a ratio that takes value close to 3 as $n \rightarrow \infty$.

Definition D.2 (Small o_p): A sequence of random vectors (including random variables), $\boldsymbol{\xi}_n$, is $o_p(1)$ if $|\boldsymbol{\xi}_n|$ converges to zero in probability. If a_n is a sequence of positive numbers, the notation $\boldsymbol{\xi}_n = o_p(a_n)$ means that $\frac{\boldsymbol{\xi}_n}{a_n} = o_p(1)$. For instance, $\sqrt{n}$ is $o(n)$ as $n \rightarrow \infty$, since $\frac{\sqrt{n}}{n} \rightarrow 0$ as $n \rightarrow \infty$.

Some important results regarding O_p and o_p are the following:

- (1) If there is a number $k > 0$ such that $E(|\boldsymbol{\xi}_n|^k)$ is bounded, then $\boldsymbol{\xi}_n = O_p(1)$; similarly, if $E(|\boldsymbol{\xi}_n|^k) \leq ca_n$, where c is a constant and a_n a sequence of positive numbers, then $\boldsymbol{\xi}_n = O_p(a_n^{1/k})$.
- (2) If there is a number $k > 0$ such that $E(|\boldsymbol{\xi}_n|^k) \rightarrow 0$, then $\boldsymbol{\xi}_n = o_p(1)$; similarly, if $E(|\boldsymbol{\xi}_n|^k) \leq ca_n$, where c is a constant and a_n a sequence of positive numbers, then $\boldsymbol{\xi}_n = o_p(b_n)$ for any sequence $b_n > 0$ such that $b_n^{-1}a_n^{1/k} \rightarrow 0$.
- (3) If there are sequences of vectors $\{\boldsymbol{\mu}_n\}$ and nonsingular matrices $\{\mathbf{A}_n\}$ such that $\mathbf{A}_n(\boldsymbol{\xi}_n - \boldsymbol{\mu}_n)$ converges in distribution, then $\boldsymbol{\xi}_n = \boldsymbol{\mu}_n + O_p(\|\mathbf{A}_n^{-1}\|)$.

Appendix E

Matrix Algebra

E.1 Matrix Differentiation

If $\mathbf{A}$ is a matrix whose elements are functions of θ , a real-valued variable, then $\frac{\partial \mathbf{A}}{\partial \theta}$ represents the matrix whose elements are the derivatives of the corresponding elements of $\mathbf{A}$ with respect to θ . For example, if

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \text{then } \frac{\partial \mathbf{A}}{\partial \theta} = \begin{pmatrix} \frac{\partial a_{11}}{\partial \theta} & \frac{\partial a_{12}}{\partial \theta} \\ \frac{\partial a_{21}}{\partial \theta} & \frac{\partial a_{22}}{\partial \theta} \end{pmatrix}$$

If $\mathbf{a} = (a_1, \dots, a_k)^T$ is a vector whose components are functions of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_l)^T$, a vector-valued variable, then $\frac{\partial \mathbf{a}}{\partial \boldsymbol{\theta}^T}$ is defined as the matrix with elements $\frac{\partial a_i}{\partial \theta_j}$, $1 \leq i \leq k, 1 \leq j \leq l$. Similarly, $\frac{\partial \mathbf{a}^T}{\partial \boldsymbol{\theta}}$ is defined as the matrix $(\frac{\partial \mathbf{a}}{\partial \boldsymbol{\theta}^T})^T$. The following are some useful results.

(1) (Inner-product) If $\mathbf{a}$, $\mathbf{b}$, and $\boldsymbol{\theta}$ are vectors, then

$$\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \mathbf{a}^T}{\partial \boldsymbol{\theta}} \right) \mathbf{b} + \left(\frac{\partial \mathbf{b}^T}{\partial \boldsymbol{\theta}} \right) \mathbf{a}.$$

(2) (Quadratic form) If $\mathbf{x}$ is a vector and $\mathbf{A}$ is a symmetric matrix, then

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A} \mathbf{x}.$$

(3) (Inverse) If the matrix $\mathbf{A}$ depends on a vector $\boldsymbol{\theta}$ and is nonsingular, then, for any component θ_i of $\boldsymbol{\theta}$,

$$\frac{\partial \mathbf{A}^{-1}}{\partial \theta_i} = -\mathbf{A}^{-1} \left(\frac{\partial \mathbf{A}}{\partial \theta_i} \right) \mathbf{A}^{-1}.$$

- (4) (Log-determinant) If the matrix $\mathbf{A}$ above is also positive definite, then, for any component θ_i of $\boldsymbol{\theta}$,

$$\frac{\partial}{\partial \theta_i} \log(|\mathbf{A}|) = \text{tr} \left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \theta_i} \right).$$

E.2 Projection

For any matrix $\mathbf{X}$, the matrix $P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}$ is called the projection matrix to $\mathcal{L}(\mathbf{X})$ (the linear space spanned by the columns of matrix $\mathbf{X}$). We assume nonsingularity for $(\mathbf{X}^T \mathbf{X})$; otherwise, $(\mathbf{X}^T \mathbf{X})^{-1}$ is replaced by the generalized inverse: $(\mathbf{X}^T \mathbf{X})^-$.

To see why we name $P_{\mathbf{X}}$ in such a way, we note that any vector in $\mathcal{L}(\mathbf{X})$ can be expressed as $\mathbf{v} = \mathbf{X}\mathbf{b}$, where $\mathbf{b}$ is a vector of the same dimension as the number of columns of $\mathbf{X}$. Then, we have $P_{\mathbf{X}}\mathbf{v} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \mathbf{X} \mathbf{b} = \mathbf{X} \mathbf{b} = \mathbf{v}$, that is, $P_{\mathbf{X}}$ does not change $\mathbf{v}$.

We define the orthogonal projection to $\mathcal{L}(\mathbf{X})$ as $P_{\mathbf{X}}^{\perp} = I - P_{\mathbf{X}}$, where I is the identity matrix. Then, for any $\mathbf{v} \in \mathcal{L}(\mathbf{X})$, we have $P_{\mathbf{X}}^{\perp} \mathbf{v} = \mathbf{v} - P_{\mathbf{X}} \mathbf{v}$. In fact, $P_{\mathbf{X}}^{\perp}$ is the projection matrix to the orthogonal space of $\mathbf{X}$, which is denoted by $\mathcal{L}(\mathbf{X})^{\perp}$.

If we define the projection of any vector $\mathbf{v}$ to $\mathcal{L}(\mathbf{X})$ as $P_{\mathbf{X}} \mathbf{v}$, then, if $\mathbf{v} \in \mathcal{L}$, the projection of $\mathbf{v}$ is itself; if $\mathbf{v} \in \mathcal{L}(\mathbf{X})^{\perp}$, the projection of $\mathbf{v}$ is zero vector.

In general, we have the orthogonal decomposition $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1 = P_{\mathbf{X}} \mathbf{v} \in \mathcal{L}(\mathbf{X})$, $\mathbf{v}_2 = P_{\mathbf{X}}^{\perp} \mathbf{v} \in \mathcal{L}(\mathbf{X})^{\perp}$ such that $\mathbf{v}_1^T \mathbf{v}_2 = \mathbf{v}^T P_{\mathbf{X}} P_{\mathbf{X}}^{\perp} \mathbf{v}$, because $P_{\mathbf{X}} P_{\mathbf{X}}^{\perp} = P_{\mathbf{X}} (I - P_{\mathbf{X}}) = P_{\mathbf{X}} - P_{\mathbf{X}}^2 = \mathbf{0}$. The last equation recalls an important property of a projection matrix; that is, any projection matrix is idempotent; that is, $P_{\mathbf{X}}^2 = P_{\mathbf{X}}$.

Appendix F

R Code

Numerous R codes were written throughout the development of this thesis. Each code served a specific purpose; however, due to their length, they have been posted to GitHub for your reference. Note, all these codes have been written for the logit link function only.

- (1) Newton-Raphson algorithm has been implemented with the line search procedure to estimate the multinomial coefficient for each category.
- (2) A code to compute the Hessian (or second derivative) for the multinomial log likelihood.
- (3) A code to compute the chi-bar square weight based on simulations and another code based on multivariate normal distribution with additional Monte Carlo steps using “ic.weights” function in the “ic.infer” R package.
- (4) A simulation code to compute the p -value for the restricted test statistics: $\bar{F}$ -test and $\bar{\chi}^2$ -test.
- (5) To find the constraint MLEs for the logit binary data, a gradient projection algorithm for the log likelihood was implemented.
- (6) To find the constraint MLEs for the logit binary data, an Iteratively Reweighted-Least Squares-Quadratic Programming method was developed.
- (7) To find the constraint MLEs for the logit multinomial data, a gradient projection algorithm for the log likelihood was implemented.
- (8) To find the unconstrained and constrained MLEs, for the multivariate GLMM, a Newton-Raphson and gradient projection algorithm was created.

To compute the chi-bar square weight, a quadratic programming approach was used as stated below.

Note F.1: For solving quadratic programming problems of the form below using R software built-in function “solve.QP”:

$$f(\boldsymbol{\theta}) = \mathbf{a}^T \boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{V}^{-1} \boldsymbol{\theta} \text{ subject to } \mathbf{A}^T \boldsymbol{\theta} \geq \boldsymbol{\theta}_0,$$

where $\mathbf{a} = -(\mathbf{V}^{-1})^T \mathbf{Z}$. If we want to minimize with the constraints $\mathbf{A}^T \boldsymbol{\theta} \leq \boldsymbol{\theta}_0$ is similar to minimize with $-(\mathbf{A}^T \boldsymbol{\theta} \geq \boldsymbol{\theta}_0)$.

All R codes mentioned in this appendix were posted to GitHub at <https://github.com/DrFSaid/CSI-CD>.

Appendix G

Distribution of Constrained MLEs for Multinomial Logit

G.1 Distribution for Case a, where all constraints are active

In this case, the data was generated assuming that both constraints are active for each of the two categories. For more information about this case, please refer to section (6.5). As you can see from the graphs, the distribution of 1000 replications for restricted MLEs appears normal except for $\tilde{\beta}_{11}$ and $\tilde{\beta}_{22}$ that are heavily skewed to the left for all sample sizes ($N = 350, 700$, and 1000).

Note: GP. β in the graph title represents $\hat{\beta}$, the restricted MLEs

Figure G.1: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case **a**

G.2 Distribution for Case b1, where at least one constraint is inactive

In case $\mathbf{b}_1$, the data was generated assuming that the first constraint is inactive for each of the two categories. For more information about this case, please refer to section (6.5). As you can see from the graphs, the distribution of 1000 replications for restricted MLEs are left and right-skewed except the distribution for $\tilde{\beta}_{21}$, where the distribution is normal for sample size ($N = 300$), and for $\tilde{\beta}_{22}$ where the distribution is normal for all sample sizes ($N = 350, 700$, and 1000).

Note: GP. β in the graph title represents $\hat{\beta}$, the restricted MLEs

Figure G.2: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b_1

Figure G.3: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b_1

Figure G.4: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b_1

G.3 Distribution for Case b2, where at least one constraint is inactive

In case **b₂**, the data was generated assuming that the second constraint is inactive for each of the two categories. For more information about this case, please refer to section (6.5). As you can see from the graphs, the distribution of 1000 replications for restricted MLEs are left and right-skewed except the distribution for $\tilde{\beta}_{02}$, where the distribution is normal for sample sizes ($N = 700$ and 1000), and for $\tilde{\beta}_{12}$ and $\tilde{\beta}_{22}$, where the distribution is normal for all sample sizes ($N = 350, 700$, and 1000).

Note: GP. β in the graph title represents $\hat{\beta}$, the restricted MLEs

Figure G.5: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b_2

Figure G.6: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b_2

Figure G.7: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case b_2

G.4 Distribution for Case **c**, where both constraints are inactive

In case **c**, the data was generated assuming that both the constraints are inactive for each of the two categories. For more information about this case, please refer to section (6.5). As you can see from the graphs, the distribution of 1000 replications for restricted MLEs are all normally distributed for all sample sizes ($N = 350, 700$, and 1000).

Note: GP. β in the graph title represents $\hat{\beta}$, the restricted MLEs

Figure G.8: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case **c**

Figure G.9: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case **c**

Figure G.10: Kernel density and histograms of constrained MLEs for $\hat{\beta}_{ij}$ for case **c**