

ADVANCES IN GLMS: FLEXIBLE APPROACHES FOR
DATA CHARACTERIZATION WITH EMPHASIS ON
HURDLE AND MULTINOMIAL MODELS

BY
NAFISA MAHMUD NASIR

A Thesis Presented to the
COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

APPLIED STATISTICS

MAY 2025

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

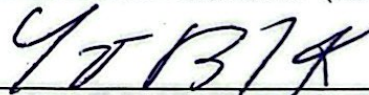
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **NAFISA MAHMUD NASIR** under the direction of her thesis adviser and approved by her thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN APPLIED STATISTICS**.

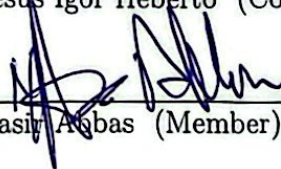
Thesis Committee



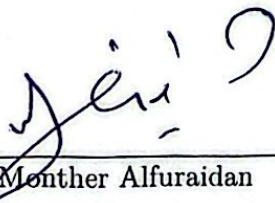
Dr. Emmanuel Osita Afuecheta (Adviser)



Dr. Jesus Igor Heberto (Co adviser)



Dr. Nasir Abbas (Member)



Prof. Menthur Alfuraiddan
Department Chairman



Prof. Suliman Al-Homidan
Dean of Graduate Studies

Date



©NAFISA MAHMUD NASIR
2025

Dedication
To my fantabulous family

ACKNOWLEDGMENTS

In the Name of Allah, the Most Gracious, the Most Merciful. All praise is due to Allah (SWT), the Nourisher, Sustainer, Evolver, Moulder, and Shaper of the world. The Lord of all worlds, who bestowed upon humanity the light of knowledge and wisdom. He created the intellect to discern truth from falsehood and granted us the ability to seek understanding. May His peace and blessings be upon the noblest of messengers, our Prophet Muhammad (SAW), the seal of the prophets, the mercy to mankind, and the guide to the straight path. May Allah also bless his family, companions, and all those who follow his guidance until the Day of Judgment. I sincerely thank Allah for extending my life and granting me the perseverance to complete this thesis successfully. I express my profound gratitude to my thesis advisor, Dr. Emmanuel Afuecheta, whose unwavering support, guidance, and encouragement were instrumental in the timely and successful completion of this work.

Without his mentorship, what now stands as reality might have remained a distant aspiration. To my co-advisor, Dr. Jesus Igor, I extend my gratitude for his insightful feedback and motivation. I also

appreciate Dr. Nasir Abbas for his constructive suggestions and discussions. I am also thankful to the dedicated faculty and staff of the Department of Mathematics for their consistent support. Your practical and insightful approach to teaching, along with your willingness to help beyond the academic realm, equipped me with the knowledge, tools, and resources essential for my academic progress. The experiences and lessons I have drawn from your expertise will undoubtedly shape both my academic and personal growth in the years to come. I extend my sincere appreciation to King Fahd University of Petroleum and Minerals and the Government of the Kingdom of Saudi Arabia for awarding me the scholarship that enabled me to pursue a master's degree, the culmination of which is this thesis. Lastly, I would like to thank my family, friends, and colleagues for their unwavering support and encouragement throughout this journey. Your words of motivation helped me navigate through challenging times. A special note of thanks goes to my beloved mother, your prayers, love, and unconditional support have borne fruit. Your daughter has made it, Alhamdulillah.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	vi
LIST OF TABLES	xi
LIST OF FIGURES	xiii
ABSTRACT (ENGLISH)	xvi
ABSTRACT (ARABIC)	xviii
CHAPTER 1 INTRODUCTION	1
1.1 Background	1
1.1.1 Multinomial Log-linear Model	3
1.1.2 Hurdle Model	4
1.1.3 Passive Smoking	7
1.2 Problem Statement	9
1.2.1 Passive Smoking and Health Risks	9
1.3 Main Goals and Objectives of this Thesis	11
1.4 Significance of the Studies	12
1.5 Layout of the Research	13
CHAPTER 2 GENERALIZED LINEAR MODEL	15
2.1 Introduction	15
2.2 Component of a GLM	20
2.2.1 Random Component	21

2.2.2	Systematic component	42
2.2.3	Link function and canonical link function	43
2.3	Parameter estimation in GLM	44
2.3.1	Estimation of the regression parameter	44
2.4	Hypothesis Testing	54
2.4.1	Analysis of deviance	54
2.5	The Poisson Approximation to the Binomial Distribution	57
2.6	Illustrative Example: Poisson Distribution	61
2.6.1	Fitting a Poisson Distribution Using the Log-Link Function . . .	62
2.6.2	Finding the Weights and Adjustments	63
2.6.3	Iteration Results	68
2.7	Discussions and Conclusion	69
CHAPTER 3 MULTINOMIAL LOG LINEAR MODEL		72
3.1	Introduction	72
3.2	Methodology	78
3.3	Data Analysis	84
3.4	Estimated Results (Fitting the Multinomial Model)	100
3.5	Discussion of Results	112
CHAPTER 4 HURDLE MODEL		187
4.1	Introduction	187
4.2	The Proposed Model	191
4.3	Methodology	194
4.3.1	Unit Inverse Gompertz Distribution (UIGD)	194
4.3.2	Comparison of Transformations	199
4.3.3	Differences Between the Two Transformations	200
4.3.4	Suitability for GLMs	202
4.3.5	Reliability and Hazard Rate Function	203
4.3.6	The p -th Ordinary Moment	209
4.3.7	Parameter Estimation	217

4.3.8	Maximum Likelihood Estimation for UIG Distribution	217
4.3.9	Model Adequacy	220
4.3.10	Unit Inverse Gumpertz Hurdle distribution (UIGHurdle)	221
4.3.11	Unit UTW Distribution	227
4.3.12	The MLE for the UIGHurdle Distribution	229
4.4	Data Analysis	231
4.5	Estimated Result and Discussion	239
CHAPTER 5	CONCLUSION AND FUTURE WORK	242
5.1	Conclusion	242
5.2	Future Work	264
APPENDIX A	PROOFS	266
APPENDIX B	SUPPLEMENTARY MATERIAL	267
REFERENCES		268
VITAE		286

LIST OF TABLES

2.1	Data for Poisson Regression Example	61
2.2	Results of iterative process	67
2.3	Iterations of Parameter Estimates	68
2.4	Final Values of β , Variance-Covariance Matrix, and Fisher Information Matrix	68
2.5	Regression Coefficients	69
2.6	Model Summary	69
3.1	Summary and Descriptive Statistics of the Entire Population by Age Group	94
3.2	Summary and Descriptive Statistics of the Male Population by Age Group	95
3.3	Summary and Descriptive Statistics of the Female Population by Age Group	95
3.4	Summary and Descriptive Statistics of the Entire Population by Region	96
3.5	Summary and Descriptive Statistics of the Male Population by Region	96
3.6	Summary and Descriptive Statistics of the Female Population by Region	97
3.7	Fitting the Multinomial Model to the Entire Population Stratified by Region	100
3.8	Continuation of Table 13	101

3.9	Fitting the Multinomial Model to the Male Population Strat- ified by Region	102
3.10	Continuation of Table 15	103
3.11	Fitting the Multinomial Model to the Female Population Stratified by Region	104
3.12	Continuation of Table 17	105
3.13	Fitting the Multinomial Model to the Entire Population Stratified by Age Group	106
3.14	Continuation of Table 19	107
3.15	Fitting the Multinomial Model to the Male Population Strat- ified by Age Group	108
3.16	Continuation of Table 21	109
3.17	Fitting the Multinomial Model to the Female Population Stratified by Age Group	110
3.18	Continuation of Table 23	111
4.1	Fitted Results for Different 0 & 1 Hurdle Models for the Dataset 1	239
4.2	Fitted Results for Different 0 & 1 Hurdle Models for Dataset 2	239
4.3	Fitted Results for Different 0 & 1 Hurdle Models for Dataset 3	240
4.4	Fitted Results Hurdle Models for the 0 and 1 Inflated Age Data.	240

LIST OF FIGURES

2.1	Plots	62
3.1	A line plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and age group	86
3.2	A line plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and administrative area	86
3.3	A bar chart illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and age group	90
3.4	A bar chart illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and administrative area	90
3.5	A box plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and age group	92
3.6	A box plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and administrative area	92
4.1	Some plots of the cdf (a) and (c), pdf(b) and (d) of the proposed distribution for selected values of β of the two transformations	207

4.2	Some plots of the reliability function (a) and (c), and hazard rate function (b) and (d) of the proposed distribution for selected values of β of the two transformations	208
4.3	Some plots of the CDF of the proposed UIGHurdle distribution for selected values of ϕ, ψ , and β of the two transformations	226
4.4	Some plots of the PDF of the proposed UIGHurdle distribution for selected values of ϕ, ψ , and β of the two transformations	226
4.5	Histogram with superimposed density curve and P–P plot of the fitted UIG-Hurdle distribution for the first transformation of the age data.	233
4.6	Histogram with superimposed density curve and P–P plot of the fitted UIG-Hurdle distribution for the second transformation of the age data.	233
4.7	Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution of the age data.	234
4.8	Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the first transformation of data set one.	234
4.9	Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the second transformation of data set one.	235
4.10	Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution of dataset one.	235
4.11	Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the first transformation of data set two.	236
4.12	Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the second transformation of data set two.	236

4.13	Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution of dataset two.	237
4.14	Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the first transformation of dataset three.	237
4.15	Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the second transformation of dataset three.	238
4.16	Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution for the first transformation of data set three.	238

THESIS ABSTRACT

NAME: NAFISA MAHMUD NASIR
TITLE OF STUDY: Advances in GLMs: Flexible Approaches for Data Characterization with Emphasis on Hurdle and Multinomial Models
MAJOR FIELD: APPLIED STATISTICS
DATE OF DEGREE: MAY 2025

Generalized Linear Models (GLMs) offer a flexible framework for modeling a diverse range of response variables by linking them to covariates through suitable link functions and probability distributions. However, their performance is contingent upon correctly specifying the underlying distribution, especially when dealing with bounded or non-standard data structures. In particular, data sets restricted to the unit interval $(0, 1)$ often necessitate the use of unit distributions to effectively capture and model meaningful variability within this bounded range. A common challenge arises when such datasets exhibit an overabundance of boundary values, specifically zeros and ones that conventional models are poorly equipped to handle. This research fundamentally seeks to develop and advance statistical modeling techniques within the framework of Generalized Linear Models (GLMs) to enhance the analysis of complex datasets that exhibit an abundance of boundary values (zeros and ones) or categorical interdependencies. To achieve this, the study proposes an extended hurdle model based on unit distributions that simultaneously accommodates excessive boundary values, thereby improving model fit and predictive

accuracy. In addition, a multinomial log-linear model is constructed to examine passive smoking exposure among people aged 10 and older in Saudi Arabia. This model investigates the associations between exposure and various sociodemographic factors, including place of exposure, age group, sex, and administrative region. By interpreting the resulting odds ratios, the study provides critical information on passive smoking patterns and supports the design of targeted public health interventions and policies. The results underscore the need for targeted interventions in high-risk settings and demographics, aligning with Vision 2030's goals for improving health.

ملخص الرسالة

الاسم: نفيسة محمود ناصر

عنوان الدراسة: التقدم في النماذج الخطية المعممة (GLMs) أساليب مرنة لوصف البيانات مع التركيز على نماذج هيردل والنماذج

اللوجستية المتعددة

التخصص: الإحصاء التطبيقي

تاريخ الدرجة العلمية: شوال، ١٤٤٦

النماذج الخطية المعممة (GLMs) توفر إطاراً مرناً لنمذجة مجموعة متنوعة من المتغيرات التابعة، من خلال ربطها بالمتغيرات المستقلة باستخدام دوال ربط مناسبة وتوزيعات احتمالية ملائمة. ومع ذلك، فإن فعالية هذه النماذج تعتمد بشكل كبير على التحديد الصحيح للتوزيع الأساسي، خاصة عند التعامل مع بيانات مقيدة أو ذات بنى غير تقليدية. على وجه التحديد، فإن مجموعات البيانات المحصورة ضمن الفاصل (0، 1) غالباً ما تتطلب استخدام توزيعات وحدوية لتمثيل التباين المعنوي داخل هذا النطاق المحدود بصورة فعالة. وتبرز مشكلة شائعة عند احتواء هذه البيانات على عدد مفرط من القيم الحدية، خصوصاً الأصفار والآحاد، وهي حالات يصعب على النماذج التقليدية التعامل معها بدقة. يتمثل الهدف الرئيسي من هذه الدراسة في تطوير وتطبيق تقنيات نمذجة إحصائية متقدمة ضمن إطار النماذج الخطية المعممة (GLMs) بهدف تحسين تحليل مجموعات البيانات المعقدة التي تتسم بوجود عدد مفرط من القيم الحدية (صفر وآحاد) أو بعلاقات تصنيفية معتمدة. ولتحقيق هذا الهدف، تقترح الدراسة نموذج هيردل موسع قائم على التوزيعات الوحدوية، بحيث يعالج في الوقت ذاته القيم الحدية الزائدة، مما يساهم في تحسين ملائمة النموذج ودقته التنبؤية. بالإضافة إلى ذلك، تم بناء نموذج لوغ-خطي متعدد الحدود لدراسة التعرض للتدخين السلبي بين الأفراد الذين تبلغ أعمارهم 10 سنوات فأكثر في المملكة العربية السعودية. ويقوم هذا النموذج بتحليل العلاقات بين التعرض للتدخين السلبي وعدد من العوامل الاجتماعية والديموغرافية، بما في ذلك مكان التعرض، الفئة العمرية، الجنس، والمنطقة الإدارية. ومن خلال تفسير نسب الأرجحية المستخلصة من النموذج، تقدم الدراسة معلومات حيوية حول أنماط التدخين السلبي، وتدعم تصميم تدخلات وسياسات صحية عامة موجهة. وتؤكد النتائج على ضرورة اعتماد تدخلات مستهدفة في البيئات والفئات السكانية عالية الخطورة، بما يتماشى مع أهداف رؤية المملكة 2030 الرامية إلى تحسين الصحة العامة.

CHAPTER 1

INTRODUCTION

1.1 Background

Ordinary Least Squares (OLS) regression is used to model the connection between one or more quantitative independent variables and a dependent variable. The method reduces the total squared errors to determine which line best fits the observation. In an ordinary least squares model, the observed variable, y , is defined as a linear function of independent variables x , reflecting a direct linear relationship. This relationship can often be visualized as a straight line. Additionally, the response variable's erroneous distribution should preferably be normal to support the model linearity assumption. Linear regression is unsuitable if no linear relationship exists between x and y . For example, a nonlinear relationship manifests itself if y increases exponentially as x increases. When applied to this type of relationship, a basic linear regression model may fail to satisfy certain fundamental assumptions of linear regression, particularly homoscedasticity. The assumption that the variability of errors in y stays fixed in any number of x is known as homoscedasticity. Our model will be affected in several ways

if this assumption is broken. Furthermore, linear regression is designed for continuous response variables y , assuming that y follows a normal distribution. Applying linear regression to a binary response variable can lead to unrealistic predictions, such as negative values, because the model inherently expects y to be continuous and distributed normally.

A generalized linear model (GLM) offers a more suitable approach [1]. GLMs allow flexibility beyond linear regression, especially when the response variable y is not normally distributed. By introducing a link function, GLMs enable a relationship between predictors and response, even when the underlying association is not strictly linear. Unlike linear regression, GLMs can accommodate various distributions for the response variable, including gamma, inverse-Gaussian, Poisson, standard, and binomial. This adaptability makes GLMs versatile for creating regression models that generalize across different contexts [1].

GLM, or the generalized linear model, is a framework that extends linear models to handle a broader range of response variable distributions. As noted above, in traditional linear regression, the dependent variable y is assumed to be normally distributed with variance τ^2 and mean α in the model. Nonnormal distributions can be included in the exponential family of distributions, incorporating the dependent variable y in GLM. Three fundamental components define the structure of a GLM:

1. **Random Component:** Identifies the distribution of the response variable y , which may include normal, binomial, Poisson distributions, etc.
2. **Systematic Component:** Defines the explanatory variables and their coefficients combined to form a linear predictor. This linear predictor is often represented as

$\eta = X\beta$, where η is a linear explanatory matrix, and β represents unknown parameters [2].

3. **Link Function:** Establishes a relationship between the mean of the dependent variable y and the explanatory η by using a specific link function (e.g., log, logit, or identity). This link function explains the relationship between the explanatory factors and the estimated value of y [3]. These elements work together to provide GLMs the flexibility of a linear predictor while modeling various data types.

1.1.1 Multinomial Log-linear Model

The multinomial log-linear model is a statistical approach to analyzing the relationships between categorical variables, often represented in contingency tables. These models provide a framework for understanding the interactions between variables by modeling the natural logarithm of expected cell counts as a function of explanatory parameters. Initially developed within contingency table analysis, the multinomial log-linear model has become a foundational tool in categorical data analysis because it can reveal complex interactions and dependencies between variables [4]. A key feature of the multinomial log-linear model is its flexibility in modeling categorical data using the exponential family of distributions. It assumes that the response variables follow a multinomial distribution, where the probabilities for each category are modeled as functions of the independent variables. The model includes main effects, which represent the impact of individual variables, and interaction terms, which capture the combined influence of two or more

variables. This ability to analyze direct effects and interactions makes the multinomial log-linear model particularly valuable in fields such as epidemiology, social sciences, and public health, where understanding complex relationships is essential [5].

1.1.2 Hurdle Model

The hurdle model has emerged as a robust statistical framework for addressing datasets with an overabundance of zero-and-one datasets. The hurdle model is a two-part statistical model designed to analyze count data with an excess of zeros, distinguishing between zero and non-zero outcomes through separate modeling processes [6]. Unlike standard Poisson or negative binomial models, which assume that all observations arise from the same underlying process, the hurdle model separates the mechanisms generating zero and positive counts, improving flexibility and interpretability in zero-inflated datasets. Hurdle models are particularly effective when excess zeros arise due to distinct behavioral or structural mechanisms rather than random variation [6]. Hurdle models can handle zero deflation (fewer zeros than expected by the data-generating process) [7]. In epidemiology, for example, they are used to model disease occurrences where many individuals may not show symptoms (zero cases). In contrast, others show varying levels of severity (positive counts) [8]. In medical studies, hurdle models have been applied to analyze hospital visit counts, medication adherence, and disease prevalence, where many individuals might never require medical care (structural zeros). In contrast, others might require multiple visits (positive counts) [9].

One notable example is the study of breast cancer patients, where a substantial proportion of individuals exhibit zero lymph node involvement, requiring a separate modeling approach to distinguish between those with and without metastatic progression [8]. Hurdle models are frequently used in passive smoking exposure studies, where many individuals may report no exposure while others experience varying levels of exposure in workplaces, public areas, and homes. Standard count models often fail to capture this pattern due to excessive zero observations, making the hurdle model more suitable for exposure assessment [10]. In economic applications, hurdle models are widely utilized to analyze consumer purchasing behaviors, where many people choose not to purchase a product (zero purchases). In contrast, others buy varying amounts (positive purchases). Similarly, they are used in criminal justice studies, where some individuals never engage in criminal activities (zero events), while others have repeat offenses [11]. Hurdle models are also effective in modeling vehicle accident frequencies, where most drivers have zero accidents, while others may have multiple incidents. The hurdle model provides more accurate risk predictions by separating the zero-event and count processes [12]. In the hurdle Poisson (HP) model, the positive count process follows a truncated Poisson distribution, which means that only observations greater than zero are modeled in the second stage [6][12]. When there is overdispersion, where the variance of the data exceeds the mean, the hurdle negative binomial (HNB) model is used, employing a truncated negative binomial distribution to model positive counts [11]. The hurdle model consists of two key components:

- **Zero-Inflation Process** – The first stage models the probability of observing

zero counts using a binary regression model, such as logistic or probit regression.

This component determines whether an observation falls into the "zero" category by estimating the likelihood of a structural zero, which arises due to fundamental constraints rather than random sampling [10].

- **Positive Count Process** – The second stage models nonzero counts using a truncated count distribution (e.g., Poisson or hostile binomial) that excludes zeros from the estimation. This step captures the variation in positive counts while accounting for overdispersion and underdispersion, which are common in real-world applications [11].

Traditional regression models often struggle with these datasets because the excessive zeros and ones violate the underlying assumptions about the distribution of the response variable. Developed as an extension of generalized linear models (GLMs), the hurdle model is particularly effective for separating binary and continuous data components, allowing a nuanced analysis of each segment independently [5]. By offering a two-part modeling framework, the hurdle model improves predictive accuracy and enhances the interpretability of results in data structures with unique features. The hurdle model splits the response variable into two stages at its core. The first stage, the binary hurdle component, models the probability of the response being zero or one, typically using a binomial distribution with a logit or probit link function [4]. This stage captures the overrepresentation of extreme values and addresses their impact on the overall dataset. The second stage focuses on continuous or discrete data beyond the hurdle, applying a truncated distribution such as Poisson or a negative bi-

nomial for nonbinary outcomes [13]. This two-part framework ensures that the binary and non-binary components are adequately accounted for, providing a comprehensive approach to modeling complex datasets.

1.1.3 Passive Smoking

Smoking remains a leading global cause of preventable morbidity and mortality, and tobacco is responsible for the deaths of half of its long-term users. Globally, more than eight million deaths occur annually due to tobacco use, including 1.2 million deaths from second-hand smoke exposure [14]. Despite extensive awareness campaigns, about 1.3 billion people worldwide continue to use tobacco products, with significant health and economic implications. In Saudi Arabia, the prevalence of smoking among adults increased from 12.2 % in 2013 to 21.4% in 2018, and many smokers consumed an average of 15 cigarettes a day [15]. This trend highlights the growing need for robust tobacco control measures. Smoking-related diseases are responsible for the deaths of five million people annually, with developing countries disproportionately affected. Alarming, 6.4 million children under the age of 18 are projected to die prematurely in adulthood due to initiating smoking during adolescence [16]. The economic burden of tobacco is equally severe, costing the United States \$167 billion annually and approximately \$160 million in tobacco purchases in Saudi Arabia each year [17]. Despite widespread awareness of its harmful effects, nicotine dependence continues to drive smoking habits. In Saudi Arabia, the cultural and religious context has played a role in antismoking measures, with King Abdulaziz declaring smoking unIslamic in 1926 [18]. These findings underscore the urgent need for more substantial tobacco control

initiatives worldwide and within Saudi Arabia.

Tobacco use is an essential contributor to noncommunicable diseases, including cancers, cardiovascular diseases, and chronic respiratory disorders such as chronic obstructive pulmonary disease (COPD). In 2018 alone, Saudi Arabia recorded more than 24,485 new cancer cases and 10,518 cancer-related deaths, with smoking as a significant factor [19]. Secondhand smoke (SHS), or environmental tobacco smoke, is particularly harmful, containing more than 7,000 chemicals, of which 70 are carcinogenic [20]. SHS increases the risk of coronary heart disease, stroke, and lung cancer by 20–30 % in non-smokers and is linked to sudden infant death syndrome in infants exposed post-birth [21]. These findings emphasize the far-reaching health consequences of tobacco use and passive smoking. In Saudi Arabia, passive smoking remains a critical problem, with 17.2% of the population exposed at home and 14.8% at work [22]. Children, in particular, are vulnerable to second-hand smoke due to their developing immune systems and face higher risks of respiratory diseases such as asthma and wheezing [23]. Alarmingly, more than one-third of university students in Saudi Arabia smoke, 36% starting the habit between the ages of 10 and 15. Parental smoking significantly impacts children’s health, further underscoring the need for targeted interventions and widespread education about the dangers of secondhand smoke.

Saudi Arabia has implemented several tobacco control initiatives in alignment with the World Health Organization’s Framework Convention on Tobacco Control (FCTC), ratified in 2005. Key measures include banning smoking in public spaces, introducing excise taxes on tobacco products, and launching antismoking campaigns [24]. The WHO MPOWER framework, adopted in 2008, has further strengthened the country’s

tobacco control efforts, focusing on monitoring tobacco use, protecting non-smokers, and offering help to quit tobacco use. However, despite these initiatives, the growing prevalence of smoking highlights the gaps in enforcement and the need for more comprehensive strategies. In conclusion, smoking and passive smoking pose significant public health and economic challenges, both globally and in Saudi Arabia. The increasing prevalence of tobacco use, coupled with its associated health risks and financial burden, requires increased efforts in tobacco control. By prioritizing education, stricter regulation, and support for quitting smoking, Saudi Arabia can address the tobacco epidemic more effectively, improving health outcomes and reducing the economic impact of smoking-related diseases.

1.2 Problem Statement

Having established the components of GLMs and the benefits of using the exponential family of distributions, let us now connect this to the problem at hand. Specifically, our objective is to apply the GLM framework to analyze the population (aged 10 years and older) exposed to passive smoking, classified by place of exposure and age group within the Kingdom.

1.2.1 Passive Smoking and Health Risks

Passive smoking, or secondhand smoke exposure, occurs when non-smokers inhale tobacco smoke from the environment. It is associated with various health risks, including respiratory diseases, cardiovascular problems, and an increased likelihood of cancer.

In Saudi Arabia, cultural norms and the prevalence of indoor smoking contribute to increased exposure, especially among children and non-smoking adults.

To assess the health impacts of passive smoking, accurate statistical models are necessary. Ignoring the effects of passive smoking can lead to:

- **Underestimated Health Risks:** Failing to account for exposure can obscure the true relationship between smoking and adverse health outcomes.
- **Misleading Policy Guidance:** Public health policies may be ineffective based on incomplete or biased analysis.
- **Inefficient Interventions:** Targeted measures may be less effective if high-risk populations are not correctly identified.

In many real-world datasets, an excessive number of zeros and ones presents a significant challenge for standard probability models, which often struggle to represent such data distributions accurately. Although hurdle models have been widely utilized to address the issue of excess zeros, there has been relatively little research on extending these models to accommodate an overabundance of both zeros and ones simultaneously. This gap in statistical modeling constrains the applicability and predictive accuracy of existing approaches, particularly in disciplines such as epidemiology, finance, and social sciences, where such data structures are commonly observed. This study aims to enhance the adaptability of hurdle models by introducing refined parameterizations that enable a more precise representation of datasets characterized by an excess of both zeros and ones. The proposed model seeks to improve the flexibility and robustness of hurdle modeling techniques by addressing this methodological

limitation, contributing to more accurate and reliable statistical inference in diverse applied research settings.

1.3 Main Goals and Objectives of this Thesis

This study aims to develop advanced statistical models for analyzing and interpreting complex data structures, particularly data sets characterized by excess zeros and ones or categorical dependencies, within the Generalized Linear Models (GLMs) framework. Specifically, it focuses on improving the hurdle and multinomial log-linear models to enhance their applicability in real-world contexts, particularly emphasizing passive smoking exposure in Saudi Arabia. By introducing innovative parameterizations and integrating robust statistical methodologies, such as bootstrapping and sensitivity analysis, this research seeks to enhance the accuracy, interpretability, and reliability of models dealing with zero-inflated and categorical data. The proposed methodological advances aim to provide more precise statistical inference, supporting evidence-based public health policies and decision-making. Furthermore, the study will explore the interaction effects between demographic and geographic factors on passive smoking exposure, contributing to a better understanding of public health risks and informing targeted intervention strategies.

The following are the objectives of this research proposal.

- Review the theoretical foundations of Generalized Linear Models (GLMs), focusing on the exponential family of distributions, link functions, and canonical links, emphasizing applications for both count and continuous data.

- To propose a multinomial log-linear model to estimate the probability distribution of passive smoking exposure across different places of exposure and age groups within the Kingdom.
- Analyze the interaction effects between demographic and geographic covariates on the probability of passive smoking exposure in varying settings and interpret the odds ratios for these interaction terms.
- Develop a hurdle model that accounts for excess zeros and ones in real datasets.

1.4 Significance of the Studies

This study is of substantial significance as, to our knowledge, it is the first to analyze passive smoking exposure in Saudi Arabia using advanced statistical models within the Generalized Linear Model (GLM) framework. Although existing research on passive smoking predominantly focuses on its health consequences, this study takes a novel approach employing hurdle and multinomial log-linear models to examine intricate exposure patterns across demographic and geographic variables. Its contribution lies in the innovative application of statistical methodologies to datasets characterized by excess zeros and categorical dependencies, addressing a gap often overlooked in traditional statistical analyses. In addition, this study provides a comprehensive understanding of passive smoking exposure within a region where indoor smoking and cultural norms contribute significantly to the prevalence of secondhand smoke. By shifting the focus to non-oil-related public health challenges, the research aligns with Saudi Arabia's Vision 2030, which prioritizes economic diversification and improved

quality of life. Integrating a multinomial log-linear model enables examining interaction effects among factors such as age, gender, place of exposure, and administrative regions. This approach extends beyond simple prevalence statistics, offering information that facilitates targeted interventions and the development of more effective public health policies.

Furthermore, the refinement of the hurdle model in this study addresses a critical gap in the literature by accurately modeling data sets with excess zeros and ones. The methodological advances introduced are relevant to public health research and extend to other fields dealing with complex data structures. By incorporating robust statistical techniques such as bootstrapping and sensitivity analysis, the study ensures the stability, reliability, and replicability of its findings, setting a benchmark for future research in this domain. The insights derived from this study are expected to benefit various stakeholders, including public health officials, policymakers, researchers, and statisticians. The findings can inform resource allocation, policy formulation, and public awareness campaigns to mitigate passive smoking exposure. Furthermore, applying advanced computational tools and statistical software, such as R, enhances the accessibility and adaptability of the methodology, facilitating its replication in other contexts. As a result, this study contributes to advancing the statistical modeling of public health research and policy development on a broader scale.

1.5 Layout of the Research

The steps of this study are described below. The first chapter provides an introduc-

tion, background study, and the aims and objectives of the study. It highlights the importance of Generalized Linear Models (GLMs) in statistical analysis. It presents the research problem, emphasizing the challenges of handling datasets with excess zeros and ones and categorical dependencies. The chapter also outlines the scope and significance of the study. Chapter 2 presents the application of the generalized linear model, focusing on the theoretical foundations of GLMs, including the exponential family of distributions, link functions, and canonical links. The review explores existing models and their limitations in dealing with datasets that exhibit unique characteristics. Chapter 3 also introduces the multinomial log-linear model to analyze categorical data, emphasizing the interaction effects between demographic and geographic covariates and passive smoking exposure in Saudi Arabia. Using data analysis techniques implemented in R, the chapter evaluates the models' performance, interprets the interaction effects, and provides insights into patterns of passive smoking. Robustness checks, including bootstrapping and sensitivity analysis, are performed to validate the stability of the findings: statistical modeling and public health. The chapter discusses the implications of these findings and suggests potential areas for future research, particularly to advance GLMs and their applications in various fields. In Chapter 4, the hurdle model is developed to address the issue of excess zeros and ones in real-world datasets. The mathematical formulation of the model is presented along with its properties. Chapter 4 delves into applying the developed models to a real-world problem.

CHAPTER 2

GENERALIZED LINEAR

MODEL

2.1 Introduction

In a traditional linear model, the expected value of y given x is expressed as a linear function of the parameters. The expected value of y given X is equal to a linear combination of x with coefficients β , and the variance of y given x is proportional to the identity matrix. When these assumptions are not met, transformations can be applied to align y with the properties of the linear model. However, finding a single transformation that simultaneously achieves a linearized mean and constant variance is often tricky. If y takes values within a limited range (e.g. binary result such as 0 or 1), modeling the expected value of y given x , which is the average or mean value of y conditional on the predictor variable(s) x as a linear function, can violate the natural constraints of the data. GLMs extend the linear model framework to address these limitations. They allow for a flexible relationship between predictors and the

response variable, making them suitable for data where the assumptions of linear models fail [25]. [26] established that a fundamental premise underlying the definition of generalized linear models is the requirement that the actual probability density of the response variable aligns with a particular distribution within the exponential family.

Generalized linear models (GLMs) were developed to extend traditional linear regression models to accommodate response variables that do not necessarily follow a normal distribution. Classical linear models assume that the response variable Y_i is normally distributed, with the expected value $E(Y_i)$ modeled as a linear function of predictors, $E(Y_i) = \mu_i = x_i^T \beta$. However, GLMs relax these assumptions, allowing the modeling of various response variables, such as counts, binary outcomes, and proportions, using appropriate distributions such as Poisson, binomial, and gamma [27]. [28] articulated that Generalized Linear Models (GLMs) serve as a highly adaptable framework for modeling the relationship between a dependent variable (response) and one or more independent variables (predictors). These models are widely used in insurance ratemaking to forecast critical metrics such as claim frequency, claim severity, pure premium, and loss ratio. Furthermore, GLMs have proven instrumental in estimating probabilities associated with binary events, including policy renewals and detecting fraudulent claims. Their versatility lies in accommodating various data distributions and link functions, making them indispensable tools for complex predictive modeling in actuarial and financial applications. [28] explained that predictors in Generalized Linear Models (GLMs) can be continuous or categorical in nature. Continuous variables are incorporated directly into the model, while categorical variables are encoded

into indicator (dummy) variables, with one level designated as the reference or base level. The choice of the base level is pivotal as it directly influences the interpretability of model coefficients and the validity of statistical significance testing. Proper selection ensures meaningful comparisons across categories and accurate inference on the effects of categorical predictors on the response variable. [2] highlights the extensive use of GLM in all disciplines due to their adaptability and versatility. In binary outcome studies, [28] explains that logistic regression, a GLM with a logit link function, is applied to model the probability of success.

Similarly, [27] emphasizes the utility of Poisson regression for count data, such as modeling traffic accident frequencies at intersections or tracking disease incidence rates in populations. For example, [27] demonstrates the value of GLM in demographic studies, which model mortality rates as a function of age and environmental factors. [28] underscores the advanced features of GLMs, such as weights and offsets, which enhance their functionality. Weights allow for differential treatment of records, ensuring that data representing larger aggregates or more significant risks exert a proportional influence on model estimation. Offsets enable the inclusion of fixed components, such as predetermined rating factors, without estimating their coefficients, thereby maintaining computational efficiency and interpretability. However, [28] also notes that GLMs face limitations, such as challenges in handling multicollinearity between predictors or capturing non-linear relationships between predictors and the response variable. Addressing these issues often requires advanced techniques or extensions to GLMs specifically designed to mitigate such constraints. [28] introduced the concept that each error structure within a generalized linear model is inherently associated with

a canonical link function, which facilitates the analytical resolution of GLM by simplifying the underlying mathematical computations. Furthermore, [28] emphasized that a valid link function must meet the essential conditions of being differentiable and monotonic, ensuring a coherent relationship between the linear predictor and the response variable.

[25] establishes that the least squares estimator, a standard tool in ordinary linear regression, also serves as the maximum-likelihood estimator when the error terms adhere to a normal distribution. Extending this principle, [25] explains that if the distribution of y belongs to the exponential family, it becomes feasible to derive maximum-likelihood estimates for the coefficients within a Generalized Linear Model (GLM). Although a numerical approximation is necessary for this estimation process, [25] notes that each iterative step can be reformulated as a weighted least squares problem. The weights are dynamically updated at each iteration, allowing the likelihood function to be optimized through the method known as *Iteratively Reweighted Least Squares* (IRLS). [27] asserts that similar to the linear model framework, certain assumptions must be satisfied when calculating a p-value using the Generalized Linear Model (GLM) approach. Violating these assumptions can compromise the model's validity by introducing biased standard errors, yielding unreliable p-values. However, [27] highlights an ongoing debate in the literature on the key assumptions that underlie GLMs and the appropriate diagnostic checks and decision-making processes required for their analysis. The calculation of Type I error (p-value) in GLMs is based on improvements in model fit, which are typically assessed using the chi-square distribution. As [27] explains, this distribution is based on assumptions of homogeneity, normality,

and independence of residuals, with deviations centered on zero. These assumptions are widely regarded as fundamental to the validity of GLM analyses. Although there is a broad consensus on the necessity of satisfying the homogeneity and independence of residuals, studies [27][29][30][5] note that meeting the assumption of independent is particularly challenging in practical applications. This underscores the importance of careful diagnostic testing and methodological rigor to ensure reliable results when employing GLM methodologies.

[5] identified the Poisson distribution as the most fundamental choice for modeling count data, positioning the Poisson regression model as a specific instance within the generalized linear model (GLM) framework. The variance of the Poisson model is equivalent to the mean, resulting in a fixed dispersion parameter (θ) set to 1. Consequently, the variance function is expressed as $V(\mu) = \mu$, reflecting this relationship. McCullagh and Nelder [29] propose that the Poisson distribution serves as the standard reference for modeling count data, analogous to how the normal distribution is regarded as the fundamental model for continuous data. John (2016) [31] highlighted that, for count data, the conditional variance of the response variable often exceeds its mean, a phenomenon termed overdispersion, which renders the Poisson distribution inadequate in such cases. This issue arises because the Poisson model assumes equality between mean and variance, implying a constant event rate and uniform probability throughout the observation period. However, real-world data frequently deviate from these assumptions, leading to heterogeneity and overdispersion, where the variance exceeds the mean. Moreover, an excess of zero observations, a common occurrence in many datasets, further challenges the applicability of the Poisson model.

[32] proposes various solutions to address the limitations of the Poisson distribution, particularly in cases of overdispersion. The negative binomial regression model is a widely used alternative, effectively managing overdispersion by adjusting standard errors and parameter estimates. The standard Poisson-based generalized linear model (GLM), which typically utilizes a log-link function, can be extended to a quasi-Poisson framework. The quasi-Poisson GLM incorporates a dispersion parameter estimated through quasi-likelihood methods, allowing it to accommodate variability that exceeds the assumptions of the Poisson distribution. Furthermore, despite not belonging to the exponential family, the negative binomial model provides a robust approach for over-dispersed count data and is estimated using maximum likelihood techniques. Furthermore, [31] proposes that data sets containing an excess of zeros beyond what the Poisson model can adequately explain are better modeled using the zero-inflated Poisson regression framework. This approach divides the data-generating process into two components: one accounts for structural zeros, while the other models the remaining counts using a Poisson-like process. By incorporating this dual mechanism, the model improves its ability to accurately fit and describe the data.

2.2 Component of a GLM

The generalized linear model (GLM) is a versatile statistical framework that generalizes linear regression to support various response variables. This is achieved by integrating three fundamental components: the random component, the systematic component, and the link function.

2.2.1 Random Component

A *random component* describes the conditional distribution of the response variable, Y_i , for the i th of n independently sampled observations based on the explanatory variables given in the model. In the original framework introduced by Nelder and Wedderburn, it is assumed that the distribution of Y_i belongs to an *exponential family* of distributions, such as the Gaussian (standard), binomial, Poisson, gamma or inverse-Gaussian families. This describes the distribution of probabilities of the response variable y . The functional form of this distribution is given by

$$f(y, \Theta, \Phi) = \exp \left[\frac{\Theta y - q(\Theta)}{p(\Phi)} + h(y, \Phi) \right] \quad (2.1)$$

where p , q , as well as h are significant functions, Φ is known as the scale component or dispersion, as well as Θ is the natural location component. In particular, the functions $q(\Theta)$ and $h(y, \Phi)$ determine the particular distribution within the exponential family.

In other words, the exponential family possesses some important properties, which include:

$$E[y] = q'(\Theta) \quad (2.2)$$

and

$$\text{Var}(y) = q''(\Theta)p(\Phi) \quad (2.3)$$

To prove (2.2) and (2.3), we proceed as follows:

$$\int_{-\infty}^{+\infty} f(y; \Theta, \Phi) dy = 1. \quad (2.4)$$

Differentiating (2.4) in terms of Θ ,

$$\int_{-\infty}^{+\infty} \frac{\partial}{\partial \Theta} f(y; \Theta, \Phi) dy = 0. \quad (2.5)$$

It is not difficult to see that from (2.1), $\log(f(y; \Theta, \Phi))$ is bound to be less tasking to differentiate:

$$\frac{\partial}{\partial \Theta} \log(f(y; \Theta, \Phi)) = \frac{\partial}{\partial \Theta} \left\{ \frac{y\Theta - q(\Theta)}{p(\Phi)} + h(y, \Phi) \right\} = \frac{y - q'(\Theta)}{p(\Phi)}. \quad (2.6)$$

Consequently, we write

$$\frac{\partial}{\partial \Theta} f(y; \Theta, \Phi) = f(y; \Theta, \Phi) \frac{\partial}{\partial \Theta} \log(f(y; \Theta, \Phi)) = f(y; \Theta, \Phi) \frac{y - q'(\Theta)}{p(\Phi)}. \quad (2.7)$$

By plugging equation (2.7) into equation (2.5), leads to

$$\int_{-\infty}^{+\infty} \frac{y - q'(\Theta)}{p(\Phi)} f(y; \Theta, \Phi) dy = 0. \quad (2.8)$$

In (2.8), we can multiply both sides by $p(\Phi)$ and the integral splits into two:

$$\int_{-\infty}^{+\infty} y f(y; \Theta, \Phi) dy - \left(\int_{-\infty}^{+\infty} f(y; \Theta, \Phi) dy \right) q'(\Theta) = 0. \quad (2.9)$$

Following the definition of the first moment, the initial term becomes $\mathbb{E}[Y]$, and the subsequent term is just $q'(\Theta)$. Hence, $\mathbb{E}[Y] = q'(\Theta)$ as required.

Next, (2.5) implies that

$$\frac{\partial}{\partial \Theta} \int_{-\infty}^{+\infty} \frac{\partial f(y; \Theta, \Phi)}{\partial \Theta} dy = \int_{-\infty}^{+\infty} \frac{\partial^2 f(y; \Theta, \Phi)}{\partial \Theta^2} dy = 0. \quad (2.10)$$

Differentiating both sides of (2.7), we have

$$\begin{aligned} \frac{\partial^2 f(y; \Theta, \Phi)}{\partial \Theta^2} &= \frac{\partial f(y; \Theta, \Phi)}{\partial \Theta} \frac{y - q'(\Theta)}{p(\Phi)} + f(y; \Theta, \Phi) \frac{-q''(\Theta)}{p(\Phi)} \\ &= f(y; \Theta, \Phi) \frac{(y - q'(\Theta))^2}{p^2(\Phi)} + f(y; \Theta, \Phi) \frac{-q''(\Theta)}{p(\Phi)} \end{aligned} \quad (2.11)$$

Now, if we Substitute (2.11) into (2.10), we have

$$\frac{1}{p^2(\Phi)} \int_{-\infty}^{+\infty} (y - q'(\Theta))^2 f(y; \Theta, \Phi) dy - \frac{q''(\Theta)}{p(\Phi)} \int_{-\infty}^{+\infty} f(y; \Theta, \Phi) dy = 0. \quad (2.12)$$

Therefore,

$$\int_{-\infty}^{+\infty} (y - q'(\Theta))^2 f(y; \Theta, \Phi) dy = q''(\Theta) p(\Phi) \int_{-\infty}^{+\infty} f(y; \Theta, \Phi) dy. \quad (2.13)$$

Recall $q'(\Theta) = \mathbb{E}[Y]$. Hence, by definition, the LHS of (2.13) is $\text{Var}(Y)$ and the RHS is $q''(\Theta)p(\Phi)$. This property provides a straightforward method for calculating the average and variability in the exponential family distribution.

Let us consider a few examples to demonstrate that many distributions follow the exponential distribution family.

1. Normal distribution

$$f(y/\alpha, \tau^2) = \frac{1}{\sqrt{2\pi\tau^2}} \exp \left\{ -\frac{(y - \alpha)^2}{2\tau^2} \right\} \quad (2.14)$$

$$= \frac{1}{(2\pi\tau^2)^{\frac{1}{2}}} \exp \left\{ -\frac{(y - \alpha)^2}{2\tau^2} \right\}$$

$$f(y \mid \alpha, \tau^2) = (2\pi\tau^2)^{-\frac{1}{2}} \exp \left\{ -\frac{(y - \alpha)^2}{2\tau^2} \right\} \quad (2.15)$$

Taking the log of equation (2.15), it yield to;

$$\log f(y \mid \alpha, \tau^2) = -\frac{1}{2} \log(2\pi\tau^2) - \frac{(y - \alpha)^2}{2\tau^2} \quad (2.16)$$

Again, taking the exponential of equation (2.16), we get;

$$f(y \mid \alpha, \tau^2) = \exp \left\{ -\frac{1}{2} \log(2\pi\tau^2) - \frac{(y - \alpha)^2}{2\tau^2} \right\} \quad (2.17)$$

By comparing equation (2.17) with,

$$f(y, \Theta, \Phi) = \exp \left\{ \frac{y\Theta - q(\Theta)}{p(\Phi)} - h(y, \Phi) \right\}$$

Now, simplifying equation (2.17), we get

$$f(y, \alpha, \tau^2) = \exp \left\{ \frac{-y^2 + 2y\alpha - \alpha^2}{2\tau^2} - \frac{1}{2} \log(2\pi\tau^2) \right\}$$

$$= \exp \left\{ \frac{2y\alpha - \alpha^2}{2\tau^2} - \frac{y^2}{2\tau^2} - \frac{1}{2} \log(2\pi\tau^2) \right\}$$

$$= \exp \left\{ \frac{y\alpha - \frac{1}{2}\alpha^2}{\tau^2} - \frac{y^2}{2\tau^2} - \frac{1}{2} \log(2\pi\tau^2) \right\}$$

where $\Theta = \alpha$, $q(\Theta) = \frac{1}{2}\alpha^2$, $p(\Phi) = \tau^2$, and

$$h(y, \Phi) = \frac{-y^2}{2\tau^2} - \frac{1}{2} \log(2\pi\tau^2) = \frac{-1}{2} \left[\frac{y^2}{\Phi} + \log(2\pi\Phi) \right]$$

$E(y)$ and $\text{Var}(y)$ are;

$$E(y) = \frac{\partial}{\partial \Theta} q(\Theta) = \frac{1}{\partial \Theta} \left[\frac{1}{2} \Theta^{-1} \right] = \frac{2}{2} \Theta$$

$$E(y) = \Theta = \alpha$$

$$\text{Var}(y) = p(\Phi) \frac{\partial^2}{\partial \Theta^2} q(\Theta) = \tau^2 \frac{\partial}{\partial \Theta} \Theta = \tau^2$$

$$\text{Var}(y) = \tau^2$$

The following is the canonical link function:

$$\Theta = \alpha$$

$$g(E(y)) = g(\alpha) = 0$$

$$g(\alpha) = 1 \text{ density}$$

Consequently, Identity serves as the canonical link function.

2. Inverse Gaussian distribution

$$f(y; \nu, \gamma) = \left(\frac{\gamma}{2\pi y^3} \right)^{1/2} \exp \left\{ -\frac{\gamma(y - \nu)^2}{2\nu^2 y} \right\} \quad (2.18)$$

Taking the logarithm of equation (2.18), we have;

$$\log f(y; \nu, \gamma) = \frac{1}{2} \left[\log \gamma - \log (2\pi y^3) - \frac{\gamma(y - \nu)^2}{2\nu^2 y} \right].$$

$$= \frac{1}{2} \log \gamma - \frac{1}{2} \log (2\pi y^3) - \frac{\gamma(y^2 - 2\nu y + \nu^2)}{2\nu^2 y}$$

$$= \frac{1}{2} \log \gamma - \frac{1}{2} \log (2\pi y^3) - \frac{\gamma y^2}{2\nu^2 y} + \frac{2\gamma \nu y}{2\nu^2 y} - \frac{\gamma \nu^2}{2\nu^2 y}$$

$$\log f(y; \nu, \gamma) = \frac{1}{2} \log \gamma - \frac{1}{2} \log (2\pi y^3) - \frac{\gamma y}{2\nu^2} + \frac{\gamma}{\nu} - \frac{\gamma}{2y} \quad (2.19)$$

Taking the exponential of equation (2.19), we get;

$$f(y; \nu, \gamma) = \exp \left\{ -\frac{\gamma}{2\nu^2} y + \frac{\gamma}{\nu} - \frac{\gamma}{2y} + \frac{1}{2} \log \gamma - \frac{1}{2} \log (2\pi y^3) \right\} \quad (2.20)$$

Comparing equation (2.20) with

$$f(y, \Theta, \Phi) = \exp \left\{ \frac{y\Theta - q(\Theta)}{p(\Phi)} + h(y, \Phi) \right\}$$

We have

$$\frac{\Theta y}{p(\Phi)} = -\frac{\gamma}{2\nu^2} y$$

Let $\Theta = \frac{1}{\mu^2}$

$$\Theta^{\frac{1}{2}} = \frac{1}{\nu} \quad \text{and} \quad \frac{1}{p(\Phi)} = -\frac{\gamma}{2}$$

$$p(\Phi) = \frac{-2}{\gamma}$$

For simplicity, we say;

$$\Phi = 1/\gamma, \quad \gamma = \frac{1}{\Phi}$$

Now, $p(\Phi) = -2\Phi$

$$f(y; \nu, \gamma) = \exp \left\{ \frac{\Theta y}{p(\Phi)} + \frac{\Theta^{1/2}}{\Phi} - \frac{1}{\Phi(2y)} + \frac{1}{2} \log \left(\frac{1}{\Phi} \right) - \frac{1}{2} \log (2\pi y^3) \right\}$$

$$f(y; \nu, \gamma) = \exp \left\{ \frac{\Theta y}{p(\Phi)} + \frac{\Theta^{\frac{1}{2}}}{\Phi} - \frac{1}{\Phi(2y)} + \frac{1}{2} \log \left(\frac{1}{\Phi} \right) - \frac{1}{2} \log (2\pi y^3) \right\}$$

$$f(y; \nu, \gamma) = \exp \left\{ \frac{\Theta y}{-2\Phi} + \frac{\Theta^{1/2}}{\Phi} - \frac{1}{2\Phi y} - \frac{1}{2} \log \Phi - \frac{1}{2} \log (2\pi y^3) \right\}$$

$$f(y; \nu, \gamma) = \exp \left\{ \frac{\Theta y - 2\Theta^{1/2}}{-2\Phi} - \frac{1}{2\Phi y} - \frac{1}{2} \log (2\pi \Phi y^3) \right\}$$

where $q(\Theta) = 2\Theta^{1/2}$, $p(\Phi) = -2\Phi$, as well as $h(y, \Phi) = -\frac{1}{2\Phi y} - \frac{1}{2} \log (2\pi \Phi y^3)$

The $E(y)$ and $\text{Var}(y)$ are;

$$E[y] = \frac{\partial}{\partial \Theta} q(\Theta) = \frac{2}{\partial \Theta} \Theta^{\frac{1}{2}}$$

$$= 2 * \frac{1}{2} \Theta^{-\frac{1}{2}}$$

$$= \Theta^{-\frac{1}{2}} = \frac{1}{\Theta^{\frac{1}{2}}} = \nu$$

$$E[y] = \nu$$

$$\text{Var}(y) = p(\Phi) \frac{\partial^2}{\partial \Theta^2} q(\Theta) = -2\Phi + \frac{\partial}{\partial \Theta} \left(\Theta^{-\frac{1}{2}} \right)$$

$$\text{Var}(y) = \Phi \Theta^{-\frac{3}{2}}$$

Recall that.

$$\Theta^{-\frac{1}{2}} = \nu$$

$$\Theta^{-\frac{3}{2}} = \nu^3$$

$$\Phi = \frac{1}{\gamma}$$

$$\text{Var}(y) = \Phi \Theta^{-\frac{3}{2}} = \frac{\nu^3}{\gamma}$$

The link function is:

$$g[E(y)] \rightarrow \Theta$$

$$g \left[\Theta^{-\frac{1}{2}} \right] \longrightarrow \Theta$$

Therefore, the canonical link function is $\frac{1}{\Theta^2}$, which means Reciprocal square.

3. Gamma distribution

$$f(y) = \frac{\beta^\alpha}{\Gamma\alpha} y^{\alpha-1} e^{-\beta y} \quad y, \beta, \alpha > 0 \quad (2.21)$$

Taking the log of equation (2.21), we have;

$$\log f(y) = \alpha \log \beta - \log \Gamma\alpha + (\alpha - 1) \log(y) - \beta y \quad (2.22)$$

Taking the exponential of equation (2.22), it yield to;

$$f(y) = \exp\{\alpha \log \beta - \log \Gamma\alpha + (\alpha - 1) \log(y) - \beta y\}$$

$$= \exp\{-\beta y + \alpha \log \beta + (\alpha - 1) \log(y) - \log \Gamma\alpha\}$$

$$= \exp\left\{\frac{-\frac{\beta y}{\alpha} + \frac{\alpha}{\alpha} \log \beta}{1/\alpha} + (\alpha - 1) \log(y) - \log \Gamma\alpha\right\}$$

$$= \exp\left\{\frac{\frac{\beta y}{\alpha} - \frac{\alpha}{\alpha} \log \beta}{-1/\alpha} + (\alpha - 1) \log(y) - \log \alpha\right\}$$

$$\text{Let } \Theta = \frac{\beta}{\alpha} \text{ and } a(\Phi) = \frac{-1}{\alpha}$$

$$\beta = \Theta\alpha, \quad \Phi = 1/\alpha$$

$$\alpha = \frac{1}{\Phi}$$

$$\beta = \frac{\Theta}{\Phi}$$

$$\log \beta = \log \Theta - \log \Phi$$

$$f(y) = \exp \left\{ \frac{\Theta y - \log \Theta}{-\Phi} + \frac{\log \Phi}{\Phi} + \left(\frac{1}{\Phi} - 1 \right) \log y - \log \Gamma \frac{1}{\Phi} \right\} \quad (2.23)$$

Comparing equation (2.23) it with,

$$f(y, \Theta, \Phi) = \exp \left\{ \frac{y\Theta - q(\Theta)}{p(\Phi)} + h(y, \Phi) \right\}$$

$$\text{Where } q(\Theta) = \log \Theta, p(\Phi) = -\Phi \text{ and } h(y, \Phi) = \frac{\log \Phi}{\Phi} + \left(\frac{1}{\Phi} - 1 \right) \log(y) - \log \Gamma \frac{1}{\Phi}$$

The mean and variance are:

$$E(y) = \frac{\partial}{\partial \Theta} q(\Theta) = \frac{\partial}{\partial \Theta} \log(\Theta) = \frac{1}{\Theta} = \Theta^{-1}$$

$$\text{Var}(y) = p(\Phi) \frac{\partial^2}{\partial \Theta^2} q(\Theta) = -\Phi \frac{\partial}{\partial \Theta} \Theta^{-1} = \frac{\Phi}{\Theta^2}$$

Recall that

$$\Theta = \frac{\beta}{\alpha}, \Phi = \frac{1}{\alpha}$$

$$E(y) = \frac{1}{\Theta} = \frac{\alpha}{\beta} = \alpha$$

$$E(y) = \frac{1}{\Theta} = \alpha$$

$$\text{Var}(y) = \frac{\Phi}{\Theta^2} = \frac{1}{\alpha} \left(\frac{\alpha^2}{\beta^2} \right) = \frac{\alpha}{\beta^2}$$

The canonical link function is:

$$\Theta = \alpha^{-1}$$

$$g(\alpha) = \Theta$$

then $g(\alpha)$ is inverse.

Therefore, the canonical link function is the inverse function.

4. Poisson distribution

$$f(y) = \frac{\kappa^y e^{-\kappa}}{y!} \quad \kappa > 0 \quad y = 0, 1, 2, \dots \quad (2.24)$$

Taking the log of equation (2.24), we have;

$$\log f(y) = y \log \kappa - \kappa - \log y! \quad (2.25)$$

Taking the exponential of equation (2.25), we have;

$$f(y) = \exp\{y \log \kappa - \kappa - \log(y!)\} \quad (2.26)$$

$$\text{let } \Theta = \log \kappa$$

$$\kappa = e^\Theta$$

Comparing equation (2.26) with

$$f(y, \Theta, \Phi) = \exp \left\{ \frac{y\Theta - q(\Theta)}{p(\Phi)} + h(y, \Phi) \right\}$$

Therefore,

$$f(y) = \exp \{y\Theta - e^\Theta - \log(y!)\}$$

$$\text{where } q(\Theta) = e^\Theta$$

$$p(\Phi) = 1, \text{ and } h(y, \Phi) = -\log(y!)$$

To find the mean and variance, we have

$$E(y) = \frac{\partial}{\partial \Theta} q(\Theta) = \frac{\partial}{\partial \Theta} e^\Theta = e^\Theta = \kappa = \alpha$$

$$\text{Var}(y) = p\Phi \frac{\partial^2}{\partial \Theta^2} q(\Theta) = 1 \frac{\partial}{\partial \Theta} e^\Theta = e^\Theta = \kappa$$

The canonical link function is:

$$g(\alpha) = \Theta$$

$$\kappa = e^{\Theta}$$

$$\log(\kappa) = \Theta$$

Therefore, $\log$ is the canonical link function of a Poisson distribution.

5. Binomial distribution

$$f(k, \kappa, \rho) = \Pr(Y = k) = \binom{\kappa}{k} \theta^k (1 - \rho)^{\kappa - k} \quad (2.27)$$

Taking $\log$ of equation (2.27), we have;

$$\log f(k, \kappa, \rho) = \log \binom{\kappa}{k} + k \log \rho + (\kappa - k) \log(1 - \rho) \quad (2.28)$$

Taking the exponential of equation (2.28), we have;

$$f(k, \kappa, \rho) = \exp \left\{ k[\log \rho - \log(1 - \rho)] + \kappa \log(1 - \rho) + \log \binom{\kappa}{k} \right\}$$

$$= \exp \left\{ k \log \left(\frac{\rho}{1-\rho} \right) + \kappa \log(1-\rho) + \log \left(\begin{matrix} \kappa \\ k \end{matrix} \right) \right\} \quad (2.29)$$

Comparing equation (2.29) with,

$$f(y, \Theta, \Phi) = \exp \left\{ \frac{\Theta y - q(\Theta)}{p(\Phi)} + h(y, \Phi) \right\}$$

$$\text{Let } \Theta = \log \left(\frac{\rho}{1-\rho} \right)$$

$$\exp(\Theta) = \frac{\rho}{1-\rho}$$

Hence, (some algebra)

$$(1-\rho)(\exp(\Theta)) = \rho$$

$$\exp(\Theta) - (\exp(\Theta))\rho = \rho$$

$$\exp(\Theta) = \rho + \rho(\exp(\Theta))$$

$$\exp(\Theta) = \rho(1 + \exp(\Theta))$$

$$\rho = \frac{\exp(\Theta)}{1 + \exp(\Theta)}$$

So,

$$1 - \rho = 1 - \frac{\exp(\Theta)}{1 + \exp(\Theta)} = \frac{1 + \exp(\Theta) - \exp(\Theta)}{1 + \exp(\Theta)}$$

$$1 - \rho = \frac{1}{1 + \exp(\Theta)}$$

$$1 - \rho = (1 + \exp(\Theta))^{-1}$$

Also,

$$\kappa \log(1 - \rho) = \kappa \log((1 + \exp(\Theta))^{-1}) = -\kappa \log(1 + \exp(\Theta))$$

$$f(k, \kappa, \rho) = \exp \left\{ k\Theta - \kappa \log(1 + \exp(\Theta)) + \log \begin{pmatrix} \kappa \\ k \end{pmatrix} \right\}$$

$$\text{Where } q(\Theta) = \kappa \log(1 + \exp(\Theta)), p(\Phi) = 1 \text{ and } h(y, \Phi) = \log \begin{pmatrix} \kappa \\ k \end{pmatrix}$$

The $E(y)$ and $\text{Var}(y)$ are:

$$E(y) = \frac{\partial}{\partial \Theta} q(\Theta) = \frac{\partial}{\partial \Theta} \kappa \log(1 + \exp(\Theta))$$

$$= \kappa \frac{\exp(\Theta)}{1 + \exp(\Theta)} = \kappa p$$

where

$$\rho = \frac{\exp(\Theta)}{1 + \exp(\Theta)}$$

$$\text{var}(y) = \rho(\Phi) \frac{\partial^2}{\partial \Theta^2} q(\Theta) = 1 \frac{\partial}{\partial \Theta} \left(\frac{\kappa \exp(\Theta)}{1 + \exp(\Theta)} \right)$$

$$\begin{aligned}
&= \kappa \frac{\exp(\Theta)}{(1 + \exp(\Theta))^2} \\
&= \kappa \left(\frac{\exp(\Theta)}{1 + \exp(\Theta)} \right) \left(\frac{1}{1 + \exp(\Theta)} \right) \\
&= \kappa \rho (1 - \rho) \text{ as expected}
\end{aligned}$$

Therefore, the canonical link function is expressed as follows.

$$\begin{aligned}
\frac{\kappa \exp(\Theta)}{1 + \exp(\Theta)} &= \alpha \\
\Rightarrow \frac{\alpha}{\kappa} &= \frac{\exp(\Theta)}{1 + \exp(\Theta)} = \rho \\
\Rightarrow \Theta &= \log \left(\frac{\rho}{1 - \rho} \right) = \log \left(\frac{\alpha/\kappa}{1 - \alpha/\kappa} \right)
\end{aligned}$$

That is a logit function.

Consequently, the logit function is the canonical link function in the binomial function.

6. Negative Binomial Distribution

$$p(Y = y) = \binom{y + r - 1}{y} (1 - p)^y p^r \quad (2.30)$$

Where r is known constant

$$0 \leq y \leq \infty$$

$$0 \leq p \leq 1$$

Taking the log of equation (2.30), we have;

$$\log f(Y = y) = \log \binom{y + r - 1}{y} + y \log(1 - p) + r \log p \quad (2.31)$$

Taking the exponential of equation (2.31), we have;

$$f(Y = y) = \exp \left\{ y \log(1 - p) + r \log p + \log \binom{y + r - 1}{y} \right\}$$

$$\text{let } \Theta = \log(1 - p)$$

$$e^{\Theta} = 1 - p$$

$$p = 1 - e^{\Theta} \quad (2.32)$$

Taking the log of equation (2.32), we have;

$$\log p = \log (1 - e^{\Theta})$$

$$f(Y = y) = \exp \left\{ \Theta y - (-r \log (1 - e^{\Theta})) + \log \binom{y + r - 1}{y} \right\}$$

$$\text{There fore, } b(\Theta) = -r \log (1 - e^{\Theta})$$

$$a(\Phi) = \Phi = 1$$

$$c(y, \Phi) = \log \binom{y + r - 1}{y}$$

The mean and variance are:

$$E[y] = \frac{\partial}{\partial \Theta} b(\Theta) = \frac{\partial}{\partial \Theta} (-r \log (1 - e^{\Theta}))$$

$$E(y) = -r \frac{(-e^{\Theta})}{1 - e^{\Theta}}$$

$$= \frac{re^{\Theta}}{1 - e^{\Theta}}, \text{ but } e^{\Theta} = (1 - p)$$

$$= \frac{r(1-p)}{p} \quad p = 1 - e^{\Theta}$$

$$E(y) = \frac{r(1-p)}{p}$$

$$\text{Var}(y) = a(\Phi) \frac{\partial^2}{\partial \Theta^2} b(\Theta)$$

$$= \frac{\partial}{\partial \Theta} \left(\frac{re^{\Theta}}{1 - e^{\Theta}} \right) = \frac{re^{\Theta}}{(1 - e^{\Theta})^2}$$

$$\text{Var}(y) = \frac{r(1-p)}{p^2}$$

Therefore, the canonical link function is expressed as:

$$g[E(y)] \Leftrightarrow \Theta$$

Let

$$\gamma = E(y) = \frac{r(1-p)}{1 - e^{\Theta}}$$

$$\gamma = \frac{re^{\Theta}}{1 - e^{\Theta}} \tag{2.33}$$

Multiply $e^{-\Theta}$ on both numerator and denominator of equation (2.33), we have;

$$= \frac{re^{\Theta} * e^{-\Theta}}{1 * e^{-\Theta} - e^{\Theta} * e^{-\Theta}}$$

$$\gamma = \frac{r}{e^{-\Theta} - 1} \quad (2.34)$$

Taking the Inverse of equation (2.34), we have;

$$\gamma^{-1} = \frac{e^{-\Theta} - 1}{r}$$

$$r\gamma^{-1} = e^{-\Theta} - 1$$

$$r\gamma^{-1} + 1 = e^{-\Theta} \quad (2.35)$$

Taking the log of equation (2.35), we have;

$$\log(r\gamma^{-1} + 1) = -\Theta$$

$$\Theta = \log(r\gamma^{-1} + 1)^{-1}$$

$$\Theta = \log \left(\frac{1}{r\gamma^{-1} + 1} \right)$$

$$\Theta = \log \left(\frac{1}{\gamma \cdot \frac{1}{r} + 1} \right)$$

$$\Theta = \log \left(\frac{1}{\frac{r}{E(y)} + 1} \right)$$

$$\frac{r}{E(y)} + 1 = \frac{r + E(y)}{E(y)}$$

$$\Theta = \log \left(\frac{E(y)}{r + E(y)} \right)$$

Therefore, the Canonical link function for Negative Binomial distribution is:

$$\Theta = \log \left(\frac{E(y)}{1 + E(y)} \right)$$

2.2.2 Systematic component

[32] established that the independent variable in a Generalized Linear Model (GLM) is defined by its systematic component. The definition of a linear mixture variables is $x_j, j = 1, 2, \dots, p$. The explanatory variables, β_i , are expressed by linearly combining unidentified variables. The quantity of the explanatory variable j^{th} is x_{ij} in every i^{th} survey, as well as $\eta_i = \sum x_j \beta_j, i = 1, 2, \dots, m$. The array syntax is $\eta = X\beta$, where η

to be known as an array of n linear independents $\eta = (\eta_1, \eta_2, \dots, \eta_n)^T$, β to be known as the matrix with m parameters $\beta = (\beta_1, \beta_2, \dots, \beta_m)^T$, and X is the (nxm) model vector expressed as

$$X = \begin{bmatrix} x_1^T \\ \cdot \\ \cdot \\ \cdot \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdot & \cdot & x_{1m} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ x_{n1} & \cdot & \cdot & x_{nm} \end{bmatrix}$$

Each item represents a distinct observation and covariate in the model vector X.

2.2.3 Link function and canonical link function

[33] states that the link function, denoted $g(\cdot)$, serves to connect the linear predictor η to the expected value of the response variable $\mathbb{E}(y)$. This link function establishes a nonlinear relationship between the mean of the dependent variable and the linear predictor coefficients, enabling the modeling of complex dependencies. The link function $g(\cdot)$ connects the linear independents to the average response variable, which means that

$$\eta_i = g_i(\alpha_i)$$

$$\eta = \mathbf{x}_i^T \boldsymbol{\beta} \quad (2.36)$$

This function needs to be differentiable and monotonic. Typically, every observation uses the same connection function. The function that converts the mean to a canonical location component of an exponential member is known as the canonical link function.

2.3 Parameter estimation in GLM

We outline the parameter estimation method within the GLM framework. Two quantities need to be calculated when fitting a GLM. The dispersion component Φ (if ambiguous) and the regression component array $\boldsymbol{\beta}$. The estimation of each of these is examined one at a time in this section.

2.3.1 Estimation of the regression parameter

There are p entries in the parameter vector $\boldsymbol{\beta}$, which can be estimated using the likelihood maximization technique. The log-likelihood value for m independent replies Y_i of equation (2.1) is given by

$$\ell(\boldsymbol{\Theta}; \mathbf{y}) = \sum_{i=1}^m \frac{y_i \Theta_i - q(\Theta_i)}{p(\Phi)} + h(y_i, \Phi) \quad (2.37)$$

This log-likelihood ratio should be maximized in relation to $\boldsymbol{\beta}$, the p score equations

$\mathbf{H}(\beta) = \partial\ell/\partial\beta = 0$ are solved for β , where the j th element of the score vector is given by

$$\frac{\partial\ell_i}{\partial\beta_j} = H_j = \sum_{i=1}^m \frac{(y_i - \alpha_i) X_{ij}}{\text{var}(Y_i)} \left(\frac{\partial\nu_i}{\partial\kappa_i} \right), j = 1, \dots, p \quad (2.38)$$

To differentiate (2.38), we employ the composite function rule

$$\partial\ell_i/\partial\beta_j = (\partial\ell_i/\partial\Theta_i) (\partial\Theta_i/\partial\nu_i) (\partial\nu_i/\partial\kappa_i) (\partial\kappa_i/\partial\beta_j)$$

as noted by [5]. In matrix notation, the score equations are defined as

$$\mathbf{H}(\beta) = \mathbf{X}^T \mathbf{Q}(\beta) \tau^{-1}(\beta) \mathbf{Z}(\beta) = \mathbf{0} \quad (2.39)$$

where $\mathbf{Q}(\beta)$ is an $m \times m$ diagonal matrix with elements $(\partial\nu_1/\partial\kappa_1, \dots, \partial\nu_m/\partial\kappa_m)$, $\tau(\beta)$ is an $m \times m$ diagonal variance-covariance matrix for $\mathbf{Y}$ with elements $\{\text{var}(Y_1), \dots, \text{var}(Y_m)\}$ and $\mathbf{Z}(\beta)$ is an $m \times 1$ vector with elements $\{y_1 - \nu_1(\beta), \dots, y_m - \nu_m(\beta)\}^T$. Notably, since these score equations are frequently non-linear and comprise p simultaneous equations with p unknowns, a numerical approach is typically used to solve them. The scoring method is the most widely applied method, essentially a Newton-Raphson method modification by which the predicted value of the second derivative is substituted for the matrix of second derivatives. The scoring method's $(t + 1)$ th iteration is provided by equation (2.1)

→ Distribution of y

$$f(y_i | \Theta_i, \Phi) = \exp \left(\frac{y_i \Theta_i - q(\Theta_i)}{p_i(\Phi)} + h(y_i, \Phi) \right)$$

where $p_i(\Phi) = \frac{\Phi}{w}$ linear predictor

$$\kappa_i = X'_{i\beta} \beta = \beta_0^\beta + X_{i1} \beta_1 + \cdots + X_{ip} \beta_p$$

⇒ Link functor

$$\alpha_i = E(y_i), \text{ and } g(\alpha_i) = \kappa_i = X'_i \beta \quad (2.40)$$

Where $g(\cdot)$ is Smooth and invertible, it is the link function taking the inverse (2.40)

$$\alpha_i = g^{-1}(\kappa_i) = g^{-1}(X_i \beta) = [q'(\Theta_i)]$$

Note that g is partly chosen to extend the range of α_i to the actual number line.

⇒ Variance

$$\text{Var}(y_i) = q''(\Theta_i) p_i(\Phi) = \underbrace{V(\alpha_i)}_{\text{variance function}} p_i(\Phi)$$

→ log likelihood function ⇒ likelihood

$$\prod_{i=1}^m f(y_i | \Theta_i, \Phi) = \exp \left\{ \sum_{i=1}^m \frac{y_i \Theta_i - q(\Theta_i)}{p_i(\Phi)} + \sum_{i=1}^m h(y_i, \Phi) \right\}$$

$\Rightarrow$ log - likelihood

$$L = \sum_{i=1}^m L_i = \sum_{i=1}^m \frac{y_i \Theta_i - q(\Theta_i)}{p_i(\Phi)} + \sum_{i=1}^m h(y_i, \Phi)$$

Maximizing the log-likelihood In terms of β Using score equations

$$\frac{\partial L_i}{\partial \beta} = \frac{\partial L_i}{\partial \Theta_i} \cdot \frac{\partial \Theta_i}{\partial \nu_i} \cdot \frac{\partial \nu_i}{\partial \kappa_i} \cdot \frac{\partial \kappa_i}{\partial \beta}$$

Where the partial derivatives are:

1.

$$\frac{\partial L_i}{\partial \Theta_i} = \frac{y_i - q'(\Theta_i)}{p_i(\Phi)} = \frac{y_i - \alpha_i}{p_i(\Phi)}$$

2.

$$\frac{\partial \kappa_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} [\beta_0 + X_i \beta_1 + \dots + X_i \beta_j] = X_{ij}$$

3.

$$\alpha_i = q'(\Theta_i),$$

then, $\frac{\partial \nu_i}{\partial \Theta_i} = q''(\Theta_i) = V(\alpha_i)$ where $q'(\Theta_i)$ is strictly increasing and invertible as well and;

$$\frac{\partial \Theta_i}{\partial \nu_i} = \frac{1}{\left(\frac{\partial \nu_i}{\partial \Theta}\right)} = \frac{1}{V(\alpha_i)}$$

4

$$g(\alpha_i) = \kappa_i \Rightarrow \alpha_i = g^{-1}(\kappa_i),$$

$$\frac{\partial \kappa_i}{\partial \nu_i} = g'(\alpha_i) \text{ then } \frac{\partial \nu_i}{\partial \kappa_i} = \frac{1}{\frac{\partial \kappa_i}{\partial \nu_i}} = \frac{1}{g'(\alpha_i)}$$

Substituting the partial derivatives into the score equations

$$\begin{aligned} \frac{\partial L_i}{\partial \beta_j} &= \frac{\partial L_i}{\partial \Theta_i} \cdot \frac{\partial \Theta_i}{\partial \nu_i} \cdot \frac{\partial \nu_i}{\partial \kappa_i} \cdot \frac{\partial \kappa_i}{\partial \beta_j} \\ &= \frac{y_i - \alpha_i}{p_i(\Phi)} \cdot \frac{1}{V(\alpha_i)} \cdot \frac{1}{g'(\alpha_i)} \cdot X_{ij} \\ \frac{\partial L_i}{\partial \beta_j} &= \sum_{i=1}^m \frac{y_i - \alpha_i}{p_i(\Phi)} \cdot \frac{1}{V(\alpha_i)} \cdot \frac{1}{g'(\alpha_i)} \cdot X_{ij} \\ \frac{\partial L_i}{\partial \beta_j} &= \sum_{i=1}^m \frac{y_i - \alpha_i}{p_i(\Phi)} \cdot \frac{1}{q''(\Theta_i)} \cdot \frac{X_{ij}}{g'(\alpha_i)} \end{aligned} \tag{2.41}$$

By definitions

$$\begin{aligned} w_i &= \frac{1}{q''(\Theta_i) g'(\alpha_i)^2} = \frac{1}{V(\alpha_i) g'(\alpha_i)^2} \\ \varepsilon_i &= X_i' \beta + g'(\alpha_i) (y_i - \alpha_i) \\ \Rightarrow \varepsilon_i - X_i' \beta &= g'(\alpha_i) (y_i - \alpha_i) \end{aligned}$$

Substituting those definitions into (2.41)

$$\begin{aligned} \frac{\partial L_i}{\partial \beta_j} &= \frac{1}{p_i(\Phi)} \sum_{i=1}^m \frac{X_{ij}}{q''(\Theta_i) g'(\alpha_i)^2} \cdot g'(\alpha_i) (y_i - \alpha_i) \\ &= \frac{1}{p_i(\Phi)} \sum_{i=1}^m w_i X_{ij} (\varepsilon_i - X_i' \beta), \quad j = 1, \dots, p \end{aligned} \tag{2.42}$$

The likelihood equation (normal equation) $\frac{\partial L_i}{\partial \beta_j} = 0$

$$\frac{1}{p_i(\Phi)} X'W(\xi - X\beta) = 0$$

$$X'W(\xi - X\beta) = 0$$

Properties of $\xi_i, i = 1, 2, \dots, m$

$$E(\xi_i) = X'\beta, \text{Var}(\xi_i) = \omega_i^{-1} p_i(\Phi)$$

i.e.

$$\begin{aligned} E(\xi_i) &= E\{X'\beta + g'(\alpha_i)(y_i - \alpha_i)\} \\ &= X'\beta + E[g'(\alpha_i)(y_i - \alpha_i)] \\ &= X'\beta + 0 \\ &= X'\beta \end{aligned}$$

$$\begin{aligned} \text{Var}(\xi_i) &= \text{Var}\{x'\beta + g'(\alpha_i)(y_i - \alpha_i)\} \\ &= \text{Var}\{X'\beta\} + \text{Var}\{g'(\alpha_i)(y_i - \alpha_i)\} \\ &= 0 + g'(\alpha_i)^2 \text{Var}(y_i - \alpha_i) \\ &= g'(\alpha_i)^2 (y_i - \alpha_i)^2 \\ &= g'(\alpha_i)^2 \text{Var}(y_i) \\ &= g'(\alpha_i)^2 q''(\Theta_i) p_i(\Phi) \\ &= g'(\alpha_i)^2 V(\alpha_i) p_i(\Phi) = \omega^{-1} p_i(\Phi) \end{aligned}$$

Recall that

$$w_i = \frac{1}{q''(\Theta_i) g(\alpha_i)^2}$$

The iteratively reweighted least squares algorithm. From equation (2.42)

$$\frac{\partial L_i}{\partial \beta_j} = \frac{1}{p_i(\Phi)} \sum_{i=1}^m \frac{X_{ij}}{q''(\Theta_i) g(\alpha_i)^2} \cdot g'(\alpha_i) (y_i - \alpha_i)$$

$$= \frac{1}{p_i(\Phi)} \sum_{i=1}^m w_i X_{ij} (\xi_i - X'_i \beta), \quad j = 1, 3 \dots p$$

$$\frac{1}{p_i(\Phi)} X'W(\xi - X\beta) = 0$$

$$X'W(\xi - X\beta) = 0 \tag{2.43}$$

This is the equation we are trying to solve

$$X'W(\xi - X\beta) = 0$$

We know that we can write equation (2.43) as

$$X'WX\beta = X'W\xi \tag{2.44}$$

Enticing to write (2.44) as

$$\beta = X'W\xi(X'WX)^{-1}$$

Where W and ξ depend on β , that is, $\beta \in W, \xi$ This is not the solution, but

$$\beta = X'W\xi(X'WX)^{-1}$$

can give us an idea of the solution for β . The Idea is set some Initial values $\beta^{(0)}$ for β , and then repeatedly calculate

$$\beta^{(a+1)} = (X'W^{(a)}X)^{-1} X'W^{(a)}\xi^{(a)}, \quad a = 0, 1, 2, 3, \dots, \quad (2.45)$$

until convergence. The right-hand side is the updating equation and $W^{(a)}$ is the matrix W and $\xi^{(a)}$ is the vector ξ . This is called the Iteratively reweighted least squares (IRLS) algorithm. To get a better idea on how $\beta^{(a)}$ become $\beta^{(a+1)}$, we assume that $e^{(a)} = \xi^{(a)} - X\beta^{(a)}$ is the working residuals at iteration $a+1$ then $\xi^{(a)} = X\beta^{(a)} + e^{(a)}$ and substituting in (2.45) We have

$$\beta^{(a+1)} = (X'W^{(a)}X)^{-1} X'W^{(a)} [X\beta^{(a)} + e^{(a)}]$$

$$\beta^{(a+1)} = \beta^{(a)} + (X'W^{(a)}X)^{-1} X'W^{(a)}e^{(a)}$$

Recall that we are solving $\frac{\partial}{\partial \beta}(\beta, \Phi, y_i) = 0$ for the MLE of β . There is already a numerical algorithm (Newton-Raphson) for doing so. The updating equation is given

by

$$\beta^{(a+1)} = \beta^{(a)} - \left(\frac{\partial^2 L_i(\beta^{(a)})}{\partial \beta \partial \beta^a} \right)^{-1} \frac{\partial}{\partial \beta} L_i(\beta^{(a)}, \Phi_1 y_i), a = 1, 2, \dots \quad (2.46)$$

We know that

$$X_{n+1} = X_n - \frac{g(X_n)}{g'(X_n)} = X_n - \frac{f'(X_n)}{f'(X_n)}$$

$$\beta^{(a+1)} = \beta^{(a)} - \left(\frac{\partial^2 L(\beta^{(a)})}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial}{\partial \beta} L(\beta^{(a)}, \Phi, y_i) \quad (2.47)$$

The matrix of second derivatives $\frac{\partial^2 L}{\partial \beta^2 \beta'}$ Consists of

$$\frac{\partial^2 L_i}{\partial \beta_j \partial \beta_k} = \left(\frac{\partial}{\partial \beta_k} \right) \left(\frac{\partial L_i}{\partial \beta_j} \right), a, k = 1, \dots, p$$

$$\frac{\partial^2 L_i}{\partial \beta_j \partial \beta_k} = \frac{1}{p_i(\Phi)} \sum_{i=1}^m X_{ij} \frac{\partial}{\partial \beta_k} (w_i g'(\alpha_i) (y_i - \alpha_i)) \quad (2.48)$$

Using (2.42) and the product rule

$$= \frac{1}{p_i(\Phi)} \sum_{i=1}^m x_{ij} (y_i - \alpha_i) \frac{\partial}{\partial \beta_k} (w_i g'(\alpha_i)) - \frac{1}{p_i(\Phi)} \sum_{i=1}^m x_j w' g'(\alpha) \nu_i \frac{\partial \alpha_i}{\partial \beta_k}(g) \quad (2.49)$$

Taking expectation, first-term vanishes since

$$E(y - \alpha_i) = 0$$

$$E \left[\frac{\partial^2 L_i}{\partial \beta_j \partial \beta_k} \right] = -\frac{1}{p_i(\Phi)} \sum_{i=1}^m X_{ij} \omega_i g'(\alpha_i) \frac{\partial \alpha_i}{\partial \beta_k}$$

$$E \left[-\frac{\partial^2 L_i}{\partial \beta_j \partial \beta_k} \right] = \frac{1}{p_i(\Phi)} \sum_{i=1}^m X_{ij} w_i g'(\alpha_i) \frac{\partial \alpha_i}{\partial \beta_k}$$

$$= \frac{1}{p_i(\Phi)} \sum_{i=1}^m w_i X_{ij} X_{ik}$$

Where $X_{ik} = g'(\alpha_i) \frac{\partial \alpha_i}{\partial \beta_k}$

The Fisher Information matrix becomes

$$L(\beta) = E \left[-\frac{\partial^2 L_i}{\partial \beta_j \partial \beta_k} \right] = \frac{1}{p_i(\Phi)} \sum_{i=1}^m w_i' X_i X_i' = \frac{1}{p_i(\Phi)} X W' X \quad (2.50)$$

The Fisher - scoring algorithm for solving $\frac{\partial}{\partial \beta} L(\beta, \Phi; y) = 0$ is obtained by replacing $\frac{\partial^2}{\partial \beta \partial \beta_k}$ in $N - R$ algorithm with its expectation $L(\beta)$. That is, we are replacing the observed information with expected fisher information. Hence, the updating equation becomes

$$\beta^{(a+1)} = \beta^{(a)} + L(\beta^{(a)})^{-1} \frac{\partial}{\partial \beta} L(\beta^{(a)}, \Phi, y_i)$$

$$\beta^{(a+1)} = \beta^{(a)} + \left(\frac{1}{p_i(\Phi)} X' W^{(a)} X \right)^{-1} \frac{1}{p_i(\Phi)} X' W^{(a)} (\xi^{(a)} - X \beta^{(a)})$$

$$\beta^{(a+1)} = \beta^{(a)} + (X'W^{(a)}X)^{-1} X'W^{(a)}e^{(a)}, \quad a = 0, 1, 2, \dots \quad (2.51)$$

2.4 Hypothesis Testing

2.4.1 Analysis of deviance

[27] state that when the response variables $Y_1, \dots, Y_N$ are independent and follow a Poisson distribution such that $Y_i \sim \text{Poisson}(\lambda_i)$, the log-likelihood function can be formally expressed as:

$$l(\beta; y) = \sum y_i \log \lambda_i - \sum \lambda_i - \sum \log y_i! \quad (2.52)$$

In the case of a saturated model, where each λ_i is distinct, the parameter vector $\beta = [\lambda_1, \dots, \lambda_N]^T$ satisfies $\hat{\lambda}_i = y_i$. Consequently, the maximum log-likelihood value becomes:

$$l(\beta_{\max}; y) = \sum y_i \log y_i - \sum y_i - \sum \log y_i!$$

For models with parameters $p < N$, the maximum likelihood estimate is used to obtain parameter estimates $\hat{\lambda}_i$, allowing the fitted values $\hat{y}_i$. Here, $E(Y_i) = \lambda_i$. The log-likelihood function for this scenario is as follows:

$$l(b; y) = \sum y_i \log \hat{y}_i - \sum \hat{y}_i - \sum \log y_i!$$

The deviance is then calculated as follows:

$$D = 2[l(\beta_{\max}; y) - l(b; y)] = 2\left[\sum y_i \log(y_i/\hat{y}_i) - \sum (y_i - \hat{y}_i)\right].$$

In most models, it can be shown that $\sum y_i = \sum \hat{y}_i$. Consequently, D simplifies as follows:

$$D = 2 \sum o_i \log \left(\frac{o_i}{e_i} \right),$$

where o_i represents the observed values y_i , and e_i denotes the estimated expected values $\hat{y}_i$.

$$D = 2 \sum [o_i \log \left(\frac{o_i}{e_i} \right) - (o_i - e_i)] \quad (2.53)$$

However, for most models $\sum o_i = \sum e_i$, the deviance simplifies to

$$D = 2 \sum [o_i \log \left(\frac{o_i}{e_i} \right)]$$

The deviance residuals are the component of D in (2.55). The standardized residuals d_i are defined as follows:

$$d_i = \text{sign}(o_i - e_i) \sqrt{2 \left[o_i \log \left(\frac{o_i}{e_i} \right) - (o_i - e_i) \right]}, \quad i = 1, \dots, N,$$

The deviance D can be expressed as:

$$D = \sum d_i^2.$$

The goodness-of-fit statistics X^2 and D are closely linked. Using the Taylor series expansion, we approximate the following:

$$o \log \left(\frac{o}{e} \right) = (o - e) + \frac{1}{2} \frac{(o - e)^2}{e} + \dots,$$

This leads to the following approximation from (2.55):

$$D = 2 \sum \left[(o_i - e_i) + \frac{1}{2} \frac{(o_i - e_i)^2}{e_i} - (o_i - e_i) \right],$$

And simplifies to:

$$D = \sum \frac{(o_i - e_i)^2}{e_i} = X^2. \tag{2.54}$$

[27] states that statistics D and X^2 are widely used to measure the goodness of fit because they are derived directly from observed data and the model. Unlike the normal distribution, these measures do not involve nuisance parameters such as σ^2 . These values can be compared with the central chi-square distribution, with degrees of freedom $N - p$, where p is the number of estimated parameters. The chi-square distribution often better approximates the sampling distribution of X^2 compared to D . Additionally, [27] notes that summary statistics available in some software include the likelihood ratio chi-square statistic and pseudo- R^2 . These statistics compare the

maximum value of the log-likelihood function for a minimal model (no covariates, i.e. $\mu_i = \log n_i + \beta_1$) with the maximum value of the log-likelihood function for a model with parameters p . The likelihood ratio chi-square statistic is calculated as:

$$C = 2[l(b) - l(b_{\min})],$$

which provides an overall test of the hypothesis that $\beta_2 = \dots = \beta_p = 0$, comparing it to the central chi-squared distribution with $p - 1$ degrees of freedom. Pseudo- R^2 is defined as:

$$R^2 = \frac{l(b) - l(b_{\min})}{-l(b_{\min})},$$

And offers an intuitive measure of fit. Other diagnostics, such as delta-betas and related statistics, can also be used for Poisson models.

2.5 The Poisson Approximation to the Binomial Distribution

The Poisson approximation to the binomial distribution is widely used in probability and statistics, particularly when analyzing rare events. It provides a simplified approach to estimating binomial probabilities when the number of trials is large and the likelihood of success is small [34]. This approximation is handy in epidemiological and public health studies, where rare occurrences, such as secondary exposure to smoking, must be effectively modeled in large populations [5]. In the context of smoking expo-

sure in various environments (e.g. cars, coffee shops, public places, homes, and workplaces), the Poisson approximation can be an effective modeling tool. When exposure events are infrequent but observed within a large population, the binomial distribution can become computationally challenging. Under these conditions, the Poisson approximation provides a more tractable model to estimate the exposure probabilities. In many settings, the likelihood of an individual being exposed to secondhand smoke (SHS) in specific locations (e.g. coffee shops or public places) is relatively low. However, when examining exposure across a large population in different geographic areas, the total number of affected individuals in each location can be more appropriately modeled using a Poisson distribution rather than a binomial distribution [4]. For example, smoking exposure in public places (e.g., restaurants and parks) is often restricted by smoking regulations. However, there are still designated smoking zones, which leads to some level of exposure risk. In large urban settings, where many daily visitors frequent public areas, the probability of encountering smoke exposure events can be effectively modeled using a Poisson distribution, simplifying calculations and providing reliable estimates for policy planning and public health interventions [35]. Using the Poisson approximation, researchers can better assess and predict smoking exposure patterns, thereby facilitating the development of effective smoking control policies and public health strategies.

It is not unexpected that the Poisson distribution can approximate the binomial distribution in some circumstances since it can be thought of as a particular form of the

other. The binomial distribution approximation using the Poisson method Given:

$$X \sim B(n, p)$$

Then X can be approximated by $Y \sim \text{Po}(\kappa)$ where $\kappa = np$ κ is the mean of the Poisson distribution.

The Poisson distribution approaches the binomial distribution more closely as the sample size (number of trials) increases. The approximation becomes precise in the limit as $n \rightarrow \infty$. [36] stated that the Poisson distribution approximates the binomial more accurately when $\kappa = np$ is smaller when the population is not infinite. The general rule is $\kappa = np < 5$ and $n > 50$. This provides a good approximation based on experience. However, the ideas regarding what constitutes a "good approximation" can vary depending on the situation. Because of this, different values for n , the number of trials in the sample, are given in some texts so that the Poisson distribution can approximate the binomial distribution. The disparate values signify the varying degrees of proximity between the two distributions. The binomial distribution, in which np is small, approaches the Poisson as n grows larger and greater. The two are so similar at some point that the differences are insignificant. However, some theorists are more circumspect than others on the required approximation degree. That is why the cited values differ.

$$\mathbf{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad (2.55)$$

We can make this look like a Poisson by letting $\kappa = np$ so that

$$\begin{aligned}
\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)!} \frac{\kappa^k}{n^k} \left(1 - \frac{\kappa}{n}\right)^{n-k} \\
&= \frac{\kappa^k}{k!} \cdot \left(1 - \frac{\kappa}{n}\right)^n \cdot \left(1 - \frac{\kappa}{n}\right)^{-k} \cdot \frac{n!}{n^k(n-k)!} \\
&= \frac{\kappa^k}{k!} \cdot \left(1 - \frac{\kappa}{n}\right)^n \cdot \left(1 - \frac{\kappa}{n}\right)^{-k} \cdot \left[\left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \cdots \left(\frac{n-k+1}{n}\right) \right]
\end{aligned}$$

If we now take $n \rightarrow \infty$ in such a way that $\kappa = np$ remains constant, then since k is fixed, we find

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\kappa}{n}\right)^n = e^{-\kappa}, \quad \lim_{n \rightarrow \infty} \left(1 - \frac{\kappa}{n}\right)^{-k} = 1$$

and

$$\lim_{n \rightarrow \infty} \left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \cdots \left(\frac{n-k+1}{n}\right) = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) = 1$$

so that

$$\lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{\kappa^k}{k!} e^{-\kappa}$$

Hence, if $X \sim \text{Bin}(n, p)$ and at least one of np and $n(1-p)$ is small so that the normal approximation does not apply, then

$$\mathbf{P}\{X = k\} \approx \mathbf{P}\{Y = k\}$$

where $Y \sim \text{Poisson}(np)$. Thus,

$$\mathbf{P}\{X = k\} \approx \frac{(np)^k}{k!} e^{-np} \quad (2.56)$$

2.6 Illustrative Example: Poisson Distribution

Fitting a Poisson distribution involves estimating the parameters β_0 and β_1 to model the relationship between the explanatory variable x_i and the response variable y_i . This is typically achieved using maximum likelihood estimation, where the model assumes that the observed counts follow a Poisson process. The canonical log link function ensures a linear relationship between the predictors and the logarithm of the expected counts while maintaining the non-negativity of μ_i .

Example

The following data set represents counts y observed at different values of a covariate x :

Table 2.1: Data for Poisson Regression Example

y_i	2	3	6	7	8	9	10	12	15
x_i	-1	-1	0	0	0	0	1	1	1

The goal is to fit a generalized linear model (GLM) to these data to model the dependence of y on x . The error distribution is assumed to follow a Poisson distribution. A $\log(y)$ plot versus x reveals an approximately linear relationship, suggesting that a log

link function is appropriate. However, y itself does not exhibit a linear dependence on x , as observed in a direct plot of y against x .

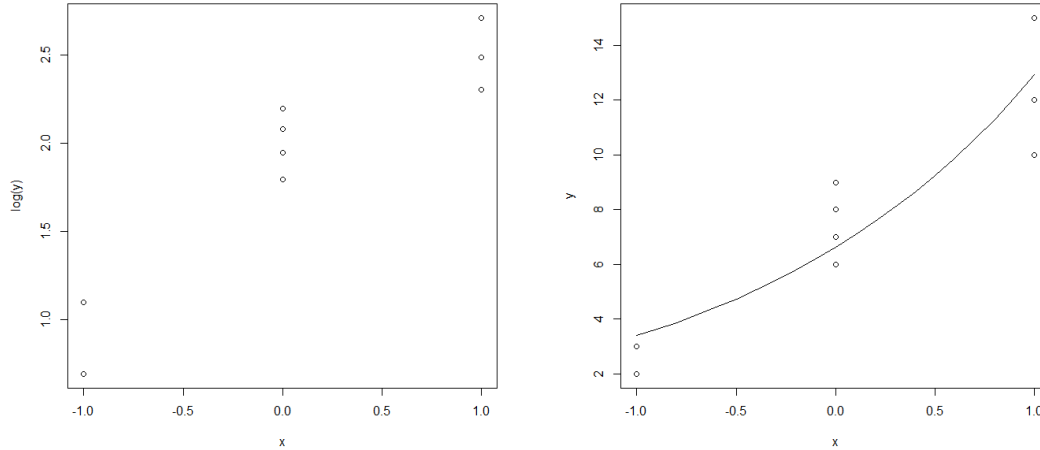


Figure 2.1: **Plots**

2.6.1 Fitting a Poisson Distribution Using the Log-Link Function

The Poisson distribution is commonly used in generalized linear models to model count data. Using the canonical log-link function, the relationship between the mean response μ_i and the explanatory variable x_i is defined as:

$$\log \mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n.$$

This log link ensures that the expected count μ_i is always positive, a critical property for count data. By exponentiating both sides, the model implies:

$$\mu_i = e^{\beta_0 + \beta_1 x_i}.$$

This shows that the relationship between x_i and μ_i is multiplicative on the original scale.

2.6.2 Finding the Weights and Adjustments

To fit this model, iterative methods such as Iteratively Reweighted Least Squares (IRLS) are employed. This requires calculating the weights w_i and adjustments ξ_i during each iteration. For the Poisson distribution:

- The link function $g(\mu) = \log \mu$ gives its derivative $g'(\mu) = 1/\mu$.
- The variance function $V(\mu) = \mu$ describes the relationship between the variance and the mean.

Using these, the weights are calculated as follows:

$$w_i = \frac{1}{V(\mu_i)[g'(\mu_i)]^2} = \frac{1}{\mu_i(\mu_i^{-1})^2} = \mu_i = e^{\beta_0 + \beta_1 x_i}.$$

The adjustment ξ_i is derived as:

$$\xi_i = \beta_0 + \beta_1 x_i + g'(\mu_i)(y_i - \mu_i),$$

which simplifies to:

$$\xi_i = \beta_0 + \beta_1 x_i + \mu_i^{-1}(y_i - \mu_i).$$

These calculations are iteratively updated until convergence, ensuring that the parameter estimates β_0 and β_1 minimize the deviance. To compute the weighted least squares matrix, we consider the following.

$$X'WX = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Expanding this product yields the following.

$$X'WX = \begin{pmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i x_i \\ \sum_{i=1}^n w_i x_i & \sum_{i=1}^n w_i x_i^2 \end{pmatrix}.$$

Substituting the weights $w_i = e^{\beta_0 + \beta_1 x_i}$, we have:

$$X'WX = \begin{pmatrix} \sum_{i=1}^n e^{\beta_0 + \beta_1 x_i} & \sum_{i=1}^n x_i e^{\beta_0 + \beta_1 x_i} \\ \sum_{i=1}^n x_i e^{\beta_0 + \beta_1 x_i} & \sum_{i=1}^n x_i^2 e^{\beta_0 + \beta_1 x_i} \end{pmatrix}.$$

Furthermore, the expression for $X'W\xi$ is given by:

$$X'W\xi = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}.$$

This simplifies to

$$X'W\xi = \begin{pmatrix} \sum_{i=1}^n w_i \xi_i \\ \sum_{i=1}^n w_i x_i \xi_i \end{pmatrix}.$$

Substituting the expression for $\xi_i = \beta_0 + \beta_1 x_i + \mu_i^{-1}(y_i - \mu_i)$, we have:

$$X'W\xi = \begin{pmatrix} \sum_{i=1}^n \mu_i (\beta_0 + \beta_1 x_i + \mu_i^{-1}(y_i - \mu_i)) \\ \sum_{i=1}^n \mu_i (\beta_0 + \beta_1 x_i + \mu_i^{-1}(y_i - \mu_i)) x_i \end{pmatrix}.$$

Expanding the terms, this becomes:

$$X'W\xi = \begin{pmatrix} \sum_{i=1}^n \mu_i (\beta_0 + \beta_1 x_i) + (y_i - \mu_i) \\ \sum_{i=1}^n \mu_i (\beta_0 + \beta_1 x_i) x_i + (y_i - \mu_i) x_i \end{pmatrix}.$$

Finally, substituting $\mu_i = e^{\beta_0 + \beta_1 x_i}$, the expression becomes:

$$X'W\xi = \begin{pmatrix} \sum_{i=1}^n (e^{\beta_0 + \beta_1 x_i} (\beta_0 + \beta_1 x_i) + y_i - e^{\beta_0 + \beta_1 x_i}) \\ \sum_{i=1}^n (e^{\beta_0 + \beta_1 x_i} (\beta_0 + \beta_1 x_i) x_i + (y_i - e^{\beta_0 + \beta_1 x_i}) x_i) \end{pmatrix}.$$

Now that we have the matrices $X'WX$ and $X'W\xi$ expressed in terms of β , we can update β iteratively. By replacing β with $\beta^{(k)}$, we compute $X'W^{(k)}X$ and $X'W^{(k)}\xi^{(k)}$ for each iteration k . These matrices are then used to calculate $\beta^{(k+1)}$ in the next step. The following updating equation defines the Iteratively Reweighted Least Squares (IRLS) algorithm:

$$\tilde{\beta}^{(k+1)} = \begin{pmatrix} \sum_{i=1}^n e^{\beta_0^{(k)} + \beta_1^{(k)} x_i} & \sum_{i=1}^n x_i e^{\beta_0^{(k)} + \beta_1^{(k)} x_i} \\ \sum_{i=1}^n x_i e^{\beta_0^{(k)} + \beta_1^{(k)} x_i} & \sum_{i=1}^n x_i^2 e^{\beta_0^{(k)} + \beta_1^{(k)} x_i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n (e^{\beta_0^{(k)} + \beta_1^{(k)} x_i} (\beta_0^{(k)} + \beta_1^{(k)} x_i) + y_i - e^{\beta_0^{(k)} + \beta_1^{(k)} x_i}) \\ \sum_{i=1}^n (e^{\beta_0^{(k)} + \beta_1^{(k)} x_i} (\beta_0^{(k)} + \beta_1^{(k)} x_i) x_i + (y_i - e^{\beta_0^{(k)} + \beta_1^{(k)} x_i}) x_i) \end{pmatrix}.$$

The algorithm starts with initial estimates $\beta_0^{(0)}$ and $\beta_1^{(0)}$. It iterates until convergence is achieved, that is when $\tilde{\beta}^{(k+1)}$ is sufficiently close to $\tilde{\beta}^{(k)}$. From the plot of $\log y$ against x in Figure 1, the initial estimates for the parameters are given as:

$$\beta_0^{(0)} = 2 \text{ (intercept)}, \quad \beta_1^{(0)} = 1 \text{ (slope)},$$

Approximately the intercept and slope of the observed linear trend. Using these initial values, we compute the following.

$$W^{(0)} = \text{diag}(2.7183, 2.7183, 7.3891, 7.3891, 7.3891, 7.3891, 20.0855, 20.0855, 20.0855),$$

$$\xi^{(0)} = (0.7358, 1.1036, 1.8120, 1.9473, 2.0827, 2.2180, 2.4979, 2.5974, 2.7468)^T,$$

$$X^T W^{(0)} X = \begin{pmatrix} 95.2494 & 54.8200 \\ 54.8200 & 65.6931 \end{pmatrix}, \quad X^T W^{(0)} \xi^{(0)} = \begin{pmatrix} 222.0694 \\ 152.5132 \end{pmatrix}.$$

In iteration 1, the updated estimates for β_0 and β_1 are calculated as:

$$\begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} = \begin{pmatrix} 95.2494 & 54.8200 \\ 54.8200 & 65.6931 \end{pmatrix}^{-1} \begin{pmatrix} 222.0694 \\ 152.5132 \end{pmatrix} = \begin{pmatrix} 1.9150 \\ 0.7235 \end{pmatrix}.$$

This iterative process continues until convergence is achieved. The subsequent table

presents the results, showing consistent parameter values in four decimal places.

Table 2.2: Results of iterative process					
k	0	1	2	3	4
$\beta_0^{(k)}$	2.0000	1.9150	1.8902	1.8892	1.8892
$\beta_1^{(k)}$	1.0000	0.7235	0.6716	0.6697	0.6697

The maximum likelihood estimates (MLE) of the parameters are as follows:

$$\hat{\beta}_0 = 1.8892 \quad \text{and} \quad \hat{\beta}_1 = 0.6697.$$

The inverse of the Fisher information matrix $I = X^T W X$ is:

$$I^{-1} = \begin{pmatrix} 0.02020 & -0.01419 \\ -0.01419 & 0.03192 \end{pmatrix},$$

which represents the variance-covariance matrix of $\hat{\beta}$. Using this, the standard errors of the estimates are:

$$\text{s.e.}(\hat{\beta}_0) = \sqrt{0.02020} = 0.1421, \quad \text{s.e.}(\hat{\beta}_1) = \sqrt{0.03192} = 0.1787.$$

The `glm` function in R produces nearly identical results for the estimated coefficients and their standard errors. For now, we focus only on the parameter estimates and standard errors.

2.6.3 Iteration Results

Table 2.3: **Iterations of Parameter Estimates**

Iteration	β_0	β_1
1	1.915027	0.7235353
2	1.890243	0.6716853
3	1.889273	0.6697881
4	1.889272	0.6697856
5	1.889272	0.6697856
6	1.889272	0.6697856
7	1.889272	0.6697856
8	1.889272	0.6697856
9	1.889272	0.6697856

Table 2.4: **Final Values of β , Variance-Covariance Matrix, and Fisher Information Matrix**

Final Beta Values	β_0	1.8892720
	β_1	0.6697856
Variance-Covariance Matrix	0.02019584	-0.01419063
	-0.01419063	0.03192892
Fisher Information Matrix	72	32.00000
	32	45.54179

Regression Output from R in Tabular Form

Table 2.5: Regression Coefficients

Coefficients	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.8893	0.1421	13.294	$< 2 \times 10^{-16}$
x	0.6698	0.1787	3.748	0.000178

Table 2.6: Model Summary

Metric	Value
Dispersion parameter for Poisson family	1
Null deviance	18.4206 on 8 degrees of freedom
Residual deviance	2.9387 on 7 degrees of freedom
AIC	41.052
Number of Fisher scoring iterations	4

2.7 Discussions and Conclusion

Applying the Poisson regression model using the Iteratively Reweighted Least Squares (IRLS) algorithm demonstrated its effectiveness in estimating the relationship between the explanatory variable x and the response variable y . Starting from initial parameter estimates of $\beta_0^{(0)} = 2$ and $\beta_1^{(0)} = 1$, the iterative process successfully converged within four iterations, achieving stable estimates of $\hat{\beta}_0 = 1.8892$ and $\hat{\beta}_1 = 0.6697$. These values align closely with the results obtained using the `glm` function in R, indicating consistency and reliability of the manual computational process.

The log link function played a critical role in the analysis, ensuring a linear relationship between the predictors and the logarithm of the expected counts while preserving the

non-negativity constraint of μ_i . This transformation was significant for modeling the count data, which inherently possesses non-negative values. The deviance measures provided insights into the model's goodness-of-fit: The null deviance, which evaluates the model without predictors, was 18.4206 with 8 degrees of freedom. The residual deviance, including the explanatory variable x , was significantly reduced to 2.9387 on 7 degrees of freedom. The substantial reduction in deviance indicates that the explanatory variable x contributes considerably to explaining the variability in the response variable y . Additionally, the Akaike Information Criterion (AIC) of 41.052 supports the model's adequacy for the data and provides a comparative measure for evaluating similar models in future studies. The variance-covariance matrix and derived standard errors further validate the model: The standard error for $\hat{\beta}_0$ was calculated as 0.1421, and for $\hat{\beta}_1$, it was 0.1787. These values indicate a high confidence level in the parameter estimates, with more minor standard errors suggesting precise and reliable estimates.

Although the Poisson regression model fits the data well, certain assumptions and limitations warrant further consideration. For instance, the model assumes equidispersion, where the variance equals the mean. Any evidence of overdispersion (variance exceeding the mean) or underdispersion could affect the validity of the results. It may require alternative models, such as the quasi-Poisson or negative binomial regression. In conclusion, the Poisson regression model with a log link function proved to be an appropriate method for analyzing the relationship between the explanatory variable x and the response variable y . The IRLS algorithm efficiently estimated the parameters, converging to stable and reliable values. The results were consistent with those

obtained from the `glm` function in R, further confirming the robustness of the models. Standard errors and deviance measures indicated that the model fits the data well, significantly reducing deviance compared to the null model. The analysis demonstrated the importance of selecting an appropriate link function and using iterative techniques for parameter estimation in generalized linear models. Future work could explore the extension of this approach to handle overdispersion, which may arise when the variance of the response variable exceeds the mean. This could involve incorporating quasi-Poisson or negative binomial models to improve flexibility. Additionally, evaluating the model's predictive performance on new data could validate its applicability further.

Finally, this study highlights the effectiveness of Poisson regression in modeling count data and underscores the importance of iterative estimation methods such as IRLS to achieve reliable parameter estimates.

CHAPTER 3

MULTINOMIAL LOG LINEAR MODEL

3.1 Introduction

As outlined in Chapter 2, the Poisson distribution is pivotal in developing and applying multinomial log-linear models, particularly to model count data in contingency tables in the context of categorical data analysis [4]. It provides a statistical foundation for estimating cell counts, often assumed to follow a Poisson distribution. The Poisson distribution offers the statistical foundation for estimating cell counts in multinomial log-linear models, extensively used to study interactions and associations among multiple variables in contingency tables. This formulation enables the model to capture complex relationships symmetrically, treating all variables equally and not developing any as a response variable [37]. [38] established the foundational importance of multinomial log-linear models in providing the statistical framework necessary to model and analyze interactions and associations within categorical data. Additionally, the Pois-

son distribution plays a crucial role in the estimation process in multinomial settings, where fixed marginal totals constrain cell counts. In such cases, although the joint distribution of cell counts follows a multinomial distribution, the Poisson framework underpins the likelihood estimation of the model parameters.

[31] highlighted the structural similarity between log-linear models for contingency tables and analysis-of-variance (ANOVA) models. These models can be fitted using generalized Poisson linear models with a logarithmic link function. Unlike traditional models that designate one variable as the response, log-linear models treat all variables symmetrically within the contingency table. This symmetric treatment ensures that the model parameters describe the associations between variables rather than the effects of explanatory variables on a predefined response variable. However, when one variable is explicitly interpreted as the response, the log-linear model transitions into a binomial or multinomial logit model.

[4] established that log-linear models, as a specialized form of generalized linear models based on the Poisson distribution, analyze the natural logarithms of expected cell counts to identify and quantify interactions between variables. Two primary types of log-linear models are the independence and saturated models. The log-linear independence model assumes that the variables in the contingency table are not associated with expected cell counts expressed as the product of their marginal probabilities. This model represents the logarithm of expected counts as an additive function of row and column effects, simplifying interpretation by isolating the independent contributions of each variable. Statistical tests such as Pearson's chi-square test and the likelihood ratio test are applied to evaluate whether the independence assumption holds, which

makes this model particularly effective for evaluating statistical independence [4]. In contrast, the log-linear saturated model is more comprehensive, incorporating interaction terms that account for all possible associations between variables. This model perfectly fits the observed data, offering a detailed representation of the relationships within the contingency table.

[39] demonstrated that the association parameters in the saturated model are directly interpretable through odds ratios, which serve as key measures of association by quantifying both the strength and direction of the variable interactions. The odds ratio provides a comparative measure of the likelihood of one outcome relative to another, offering a meaningful interpretation of dependency structures within contingency tables. Notably, when the interaction parameters in the saturated model are zero, the model simplifies to the independence model, signifying the absence of associations between variables. In this case, expected cell counts can be expressed solely as the product of their marginal probabilities, confirming statistical independence between categorical variables. These models form a robust analytical framework for categorical data and have broad applications in medicine, social sciences, and public health [40]. Their ability to reveal associations and interactions makes them valuable for understanding relationships in multivariate categorical data. The odds of an event are defined as the ratio of the probability that the event occurs to the probability that it does not happen, mathematically expressed as:

$$\text{Odds} = \frac{p}{1 - p}.$$

For example, if the probability of success is 0.75, the odds are 3, indicating that the event is three times more likely to occur than not. Building on this concept, the odds ratio (OR) compares the odds of an event in one group with the odds in another, serving as a key measure to quantify the strength and direction of the association between two categorical variables. Commonly utilized in contingency table analysis and logistic regression, the OR is calculated as:

$$\text{OR} = \frac{\text{Odds in Group 1}}{\text{Odds in Group 2}},$$

or, in a 2×2 contingency table, as:

	Outcome Present (A)	Outcome Absent
Exposed	a	b
Unexposed	c	d

$$\text{Odds Ratio (OR)} = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} = \frac{ad}{bc}$$

Here, a, b, c , and d represent the respective cell counts. [4] established that the odds ratio is a fundamental measure of association in statistical modeling. An odds ratio of 1 signifies that there is no association between the variables, indicating that the occurrence of one event does not affect the probability of the other. In contrast, odds ratios greater than 1 suggest a positive association, which means that the likelihood of the associated event increases as one variable increases. In contrast, odds ratios less than 1 indicate a negative association, implying that an increase in one variable

corresponds to a decrease in the likelihood of another event. This interpretability makes the odds ratio crucial for assessing relationships in contingency tables and regression models.

[4] highlighted several notable properties of the odds ratio that enhance its applicability in statistical analysis. One key property is its symmetry, meaning that interchanging the groups being compared does not alter its value. Furthermore, the odds ratio remains independent of marginal totals in a contingency table, making it a robust and reliable measure for evaluating associations in cross-tabulated data. Furthermore, the logarithmic transformation of the odds ratio, expressed as $\log(\text{OR})$ is frequently utilized in statistical models due to its symmetric distribution and ease of interpretation. This transformation simplifies analysis, particularly in logistic regression, where the odds ratio is critical in quantifying the effect of predictor variables on the odds of an outcome. Unlike the relative risk, which compares probabilities directly, the odds ratio approximates relative risk only when the event's probability is small. Given these advantages, the odds ratio remains indispensable in categorical data analysis. It is widely applied in epidemiology, social sciences, and public health research to assess the relationships between variables and inform decision-making.

In log-linear models, the requirement for the Poisson mean (μ) to remain positive makes the additive model, expressed as

$$\mu = \sum \beta_j x_j,$$

Both scientifically and logically inappropriate, particularly for extrapolation, as it can

lead to negative values under specific parameters or combinations of covariates. To resolve this, log-linear models adopt a multiplicative approach where

$$\mu = \exp(\eta) \quad \text{and} \quad \eta = \sum \beta_j x_j,$$

[5] emphasized the importance of ensuring that μ remains strictly positive regardless of parameter or covariate values. This structural formulation is particularly advantageous as it prevents the possibility of negative values, which are scientifically and logically inconsistent in count-based models. This framework facilitates the incorporation of factorial design and regression modeling concepts, allowing model parameters to be interpreted as contrasts of logarithmic frequencies. Since working directly with log-transformed values can be less intuitive, these contrasts are typically back-transformed to the original frequency scale for improved clarity and interpretation. This transformation enables results to be expressed in terms of multiplicative effects, which are more intuitive and applicable in practical statistical analyses.

[3] established that the model goodness of fit is typically evaluated using likelihood-based methods, including the likelihood ratio test and Pearson's chi-square statistic in the multinomial log-linear framework. These statistical measures compare the observed frequencies in the contingency tables with the expected frequencies predicted under the null hypothesis of independence. Large discrepancies between observed and expected frequencies indicate significant associations between variables. The rigorous nature of these tests ensures the statistical validity and reliability of conclusions drawn from multinomial log-linear analyses, particularly in studies involving large and

complex datasets. [33] highlighted the broad applicability of multinomial log-linear models in multiple disciplines. In public health research, these models are widely used to analyze disease risk factors by assessing relationships between demographic variables and health outcomes. For example, in passive smoking studies, multinomial log-linear models facilitate the investigation of how age group, place of exposure, and geographic region interact to influence exposure probabilities. By explicitly modeling interactions among categorical variables, these models provide granular insights essential for developing effective public health interventions and policies.

[2] underscored that recent advancements in computational techniques and statistical software have significantly improved the efficiency and accessibility of multinomial log-linear models. Modern tools such as R's `nnet` package enable computationally efficient parameter estimation and model validation, even for highly complex datasets. Furthermore, integrating robust statistical methods, including bootstrapping and sensitivity analysis, has improved the stability and reliability of parameter estimates and predictions. These methodological advancements have cemented multinomial log-linear models as an indispensable tool in contemporary statistical analysis, fostering progress in research, decision-making, and policy formulation across diverse fields.

3.2 Methodology

The log-linear model is a statistical technique used to analyze the associations between categorical variables in contingency tables. As a linear function of the parameters corresponding to each category variable and their interactions, it predicts the

natural logarithm of the expected cell counts. According to [4], log-linear models are frequently utilized to evaluate the interactions between variables or determine their independence. Suppose that we categorize and/or use two criteria to categorize a random sample of n subjects A and B , for example, in our case, age groups and the place of exposure to smoking or the administrative area and the place of exposure to smoking. Let y_{ij} be the number of subjects in each category i of A and category j of B , where $i = 1, \dots, I$ and $j = 1, \dots, J$. We have the following:

$$\sum_{i=1}^I \sum_{j=1}^J y_{ij} = m.$$

In this case, our data $\{y_{ij}\}$ form a 2-D contingency table as follows:

	B_1	B_2	$\cdots$	B_J	Sum
A_1	y_{11}	y_{12}	$\cdots$	y_{1J}	$y_{1\cdot}$
A_2	y_{21}	y_{22}	$\cdots$	y_{2J}	$y_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}	$y_{I\cdot}$
Sum	$y_{\cdot 1}$	$y_{\cdot 2}$	$\cdots$	$y_{\cdot J}$	m

where the column sum $y_{\cdot j} = \sum_{i=1}^I y_{ij}$, row sum $y_{i\cdot} = \sum_{j=1}^J y_{ij}$, for $i = 1, \dots, I$ and $j = 1, \dots, J$. The total is:

$$y_{\cdot\cdot} = \sum_{i=1}^I \sum_{j=1}^J y_{ij} = m.$$

The intriguing question is whether there is a connection between age groups and the

place of exposure to smoking or between the administrative area and the place of exposure to smoking. Thus, the distribution of $\{Y_{ij}\}$ is assumed to be multinomial:

$$P(Y_{11} = y_{11}, \dots, Y_{IJ} = y_{IJ}) = \frac{m!}{y_{11}! \dots y_{IJ}!} \pi_{11}^{y_{11}} \pi_{IJ}^{y_{IJ}}, \quad y_{11}, \dots, y_{IJ} \geq 0, \quad y_{11} + \dots + y_{IJ} = m,$$

where π_{ij} are designated to be the cell probabilities. For example, π_{11} denotes the likelihood that a randomly chosen research subject will find itself in row 1 and column 1. As a result, the marginal chances are given by $\pi_{i.} = \sum_j \pi_{ij}$ and $\pi_{.j} = \sum_i \pi_{ij}$, for $i = 1, \dots, I, j = 1, \dots, J$. Now, to test the hypothesis of no association (independence) between subjects A and B , we can write the null hypothesis as follows:

$$H_0 : \pi_{ij} = \pi_{i.} \pi_{.j}, \quad \text{for all } i, j.$$

The alternative hypothesis is the following.

$$H_1 : H_0 \text{ is not true.}$$

It is not difficult to see the likelihood ratio test as obtained in statistical inference to be approximately the χ^2 test statistic:

$$X^2 = \sum_{ij} \frac{(y_{ij} - m\hat{\pi}_{i.}\hat{\pi}_{.j})^2}{m\hat{\pi}_{i.}\hat{\pi}_{.j}} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where $\hat{\pi}_{i.} = \frac{y_{i.}}{m}$, $\hat{\pi}_{.j} = \frac{y_{.j}}{m}$, $O_{ij} = y_{ij}$ represents the observed frequency and $E_{ij} = m\hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{y_{i.}y_{.j}}{m}$ (row total $\times$ column total / grand total), for $i = 1, \dots, I, j = 1, \dots, J$

is the expected frequency.

If H_0 is true, the distribution of X^2 is approximately χ^2 with $(I - 1)(J - 1)$ degrees of freedom. In particular, a large value of X^2 is taken as significant evidence against H_0 , and we reject H_0 if and only if:

$$X^2 > \chi_{\alpha; (I-1)(J-1)}^2,$$

The critical value in the χ^2 distribution with $(I - 1)(J - 1)$ DF. Similarly, we can reject H_0 if the value of P associated with the value of X^2 is less than the level α .

Similarly, a Poisson model can describe the data of the contingency table and test the necessary hypothesis regarding the association between the row and column classifications. In this case:

$$Y_{ij} = \alpha_{ij} + \varepsilon_{ij} \sim \text{Pois}(\alpha_{ij}) \text{ independent, } (\alpha_{ij} = m\pi_{ij}, \text{ for all } i, j), \quad (3.1)$$

where

1. Y_{ij} : The observed count for the $(ij)^{\text{th}}$ category.
2. α_{ij} : The expected mean of the Poisson distribution for the $(ij)^{\text{th}}$ category, which is modeled as $m\pi_{ij}$, where:

- m : Total number of observations (a fixed sum of counts).
- π_{ij} : Probability associated with the $(ij)^{\text{th}}$ category.

3. ε_{ij} : Random deviation from the expected mean.
4. $Y_{ij} \sim \text{Pois}(\alpha_{ij})$: Each count follows a Poisson distribution with mean α_{ij} , and the counts between cells are independent.

The log-linear representation of the Poisson means is given by;

$$\log(\alpha_{ij}) = \alpha + \alpha_i + \beta_j + \gamma_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (3.2)$$

The log-linear formulation specifies that α_{ij} is modeled in terms of:

- α : An overall mean term.
- α_i : A main effect for the i^{th} row (for example, a variable effect at the row level).
- β_j : A main effect for the j^{th} column (for example, a column-level variable effect).
- γ_{ij} : An interaction term capturing the combined effects of the i^{th} and j^{th} levels.

In this class of models, it is assumed that the sum $Y_{..}$ is distributed according to the Poisson. That is, $Y_{..} = Y_{11} + \dots + Y_{IJ} \sim \text{Pois}(\alpha_{..})$, where $\alpha_{..} = \alpha_{11} + \dots + \alpha_{IJ} = m(\pi_{11} + \dots + \pi_{IJ}) = m$. The total count $Y_{..}$ in categories is distributed as $\text{Pois}(\alpha_{..})$, where $\alpha_{..}$ is the sum of the means of the Poisson distributions in all cells. The constraint $(\pi_{11} + \dots + \pi_{IJ}) = 1$ ensures that $\alpha_{..} = m$, making $Y_{..}$ a fixed constant. Hence, the conditional distribution of $\{Y_{ij}\}$ given $Y_{..} = m$ is multinomial with the functional form:

$$P(Y_{11} = y_{11}, \dots, Y_{IJ} = y_{IJ} \mid Y_{..} = m) = \frac{P(Y_{11} = y_{11}, \dots, Y_{IJ} = y_{IJ})}{P(Y_{..} = m)}$$

The joint probability of Poisson counts is given by

$$P(Y_{..} = y_{..}, \dots, Y_{IJ} = y_{IJ}) = \prod_{ij} \frac{(m\pi_{ij})^{y_{ij}} \exp(-m\pi_{ij})}{y_{ij}!}$$

This is the product of independent Poisson probabilities for all $(i, j)^{\text{th}}$ cells. The marginal probability of total counts is

$$P(Y_{..} = m) = \frac{m^m \exp(-m)}{m!}$$

This represents the marginal probability of observing a total count m , derived from the properties of the Poisson distribution. Using the definition of conditional probability,

$$P(Y_{ij} = y_{ij}, \dots, Y_{IJ} = y_{IJ} \mid Y_{..} = m) = \frac{P(Y_{11} = y_{11}, \dots, Y_{IJ} = y_{IJ})}{P(Y_{..} = m)} \quad (3.3)$$

Substituting the expressions for the numerator (joint Poisson probability) and the denominator (marginal total likelihood) of (3.3)

$$P(Y_{ij} = y_{ij}, \dots, Y_{IJ} = y_{IJ} \mid Y_{..} = m) = \frac{\prod_{ij} \frac{(m\pi_{ij})^{y_{ij}} \exp(-m\pi_{ij})}{y_{ij}!}}{\frac{m^m \exp(-m)}{m!}}$$

Cancel $\exp(-m)$ in the numerator and denominator. Using the fact that $\sum_{ij} y_{ij} = m$,

$\sum_{ij} \pi_{ij} = 1$, and simplify factorial terms.

$$= \prod_{ij} \frac{m^{y_{ij}} (\pi_{ij})^{y_{ij}}}{y_{ij}!} \cdot \frac{m!}{m^m \exp(-m) \exp(m)}.$$

Final expression:

$$P(Y_{11} = y_{11}, \dots, Y_{IJ} = y_{IJ} \mid Y_{..} = m) = \frac{m!}{m^m} \prod_{ij} \frac{(m\pi_{ij})^{y_{ij}}}{y_{ij}!}.$$

$$P(Y_{11} = y_{11}, \dots, Y_{IJ} = y_{IJ} \mid Y_{..} = m) = \frac{m!}{\prod_{ij} y_{ij}!} \prod_{ij} \pi_{ij}^{y_{ij}}, \quad \text{if } y_{11} + \dots + y_{IJ} = m \quad (3.4)$$

This is the probability mass function of a multinomial distribution with m trials and cell probabilities π_{ij} . The null hypothesis $H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$ for all i, j becomes:

$$\log(\alpha_{ij}) = \log(m\pi_{i.}\pi_{.j}) = \alpha + \alpha_i + \beta_j, \quad \text{i.e., } \gamma_{ij} = 0 \text{ for all } i, j.$$

Consequently, the test of independence and/or no association becomes a test of no interaction. Now, since the denominator of (3.4), $\frac{n^n e^{-n}}{n!}$ is independent of any unknown parameters, it means that the MLE in the Poisson model is similar to that in the multinomial model.

3.3 Data Analysis

The dataset investigated in this research was downloaded from the Saudi National Statistics Authority [<https://www.stats.gov.sa/en/statistics-overview>]. Specifically,

it comprises the National Household Health Survey, which took place in 2017 and aims to create a comprehensive perspective on healthcare status in the Kingdom. The survey assesses various indicators, including insurance coverage, disease prevalence, lifestyle behaviors, and smoking habits. A random sample of 1,334 households was drawn in all regions of the Kingdom. It is beyond the scope of this work to provide a detailed explanation of this national survey. Readers interested in digging deeper can access the following reference [<https://www.stats.gov.sa/sites/default/files/household-health-survey-2017-1.pdf>]. The operations carried out with the mentioned dataset are explained in the following lines. This data set provides comprehensive information on the entire Saudi population and separate data for male and female populations aged 10 years and older. It is organized by administrative regions within Saudi Arabia and is further divided into age groups ranging from 10 to 65 years of age or older. For each region and age group, the data set captures the number of people exposed to smoking in five places: cars, coffee shops, homes, public areas, and workplaces.

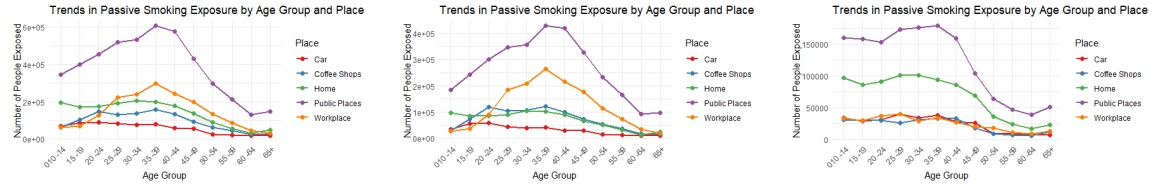


Figure 3.1: A line plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and age group

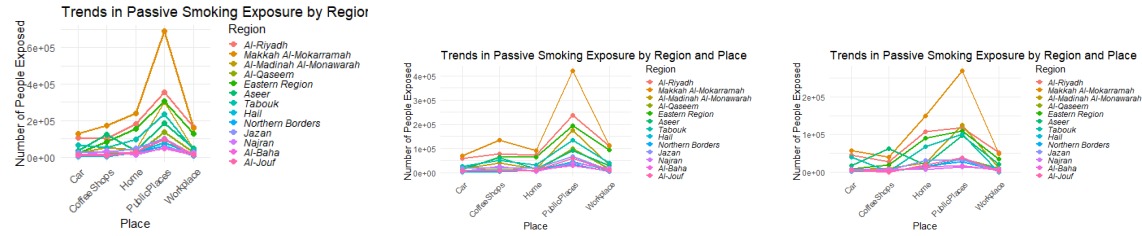


Figure 3.2: A line plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and administrative area

Figure 2a represents the exposure of the entire population to smoking in different places. The data indicates that exposure peaks around the 35-39 age group, particularly in public areas and workplaces. After age 40, there is a consistent decline in exposure in all places, suggesting that smoking exposure decreases with age. This decline may be attributed to lifestyle changes, reduced public participation, or health-related considerations as people transition to middle and older age. Furthermore, younger individuals (aged 10-24 years) exhibit relatively lower levels of smoking exposure compared to middle-aged groups (25-44 years). This difference could reflect variations in the environments visited by these age groups or the smoking regulations in the places they visit. In general, the data underscore the influence of age and social behavior on smoking exposure patterns. Figure 2b illustrates the passive smoking exposure patterns among the male population in various places. Public places

and workplaces emerge as the dominant sources of exposure, particularly for individuals in the 35-39 age range. In contrast, home exposure remains steady across all age groups, indicating a persistent risk of passive smoking at home regardless of age. Coffee shops and cars show lower levels of exposure in all ages, with cars being the least significant source of passive smoking. In particular, there is a sharp decline in exposure to smoking in public places and workplaces after the age of 40. This trend likely corresponds to life changes, such as retirement or reduced participation in public spaces as individuals age. The data underscore the importance of public places and workplaces as the main contributors to passive smoking exposure, particularly for younger and middle-aged adults. These findings provide valuable information for public health interventions, emphasizing the need to focus efforts on reducing exposure in these environments to mitigate health risks. Figure 2c illustrates the passive smoking exposure patterns among the female population in various places. The 30-39 - year age group emerges as the most exposed demographic, particularly in public places, homes, and workplaces. In contrast, older age groups (ranging from 40-44) experience a noticeable sharp decline in passive smoking exposure, especially in public places and workplaces. This trend may reflect lifestyle changes, reduced participation in public activities, or health-related factors as individuals age. Public places consistently represent the most significant source of passive smoking exposure for females, whereas cars and coffee shops contribute the least across all age ranges. Data provide a detailed understanding of how exposure to passive smoking varies by place and age group, with younger and middle-aged adults experiencing greater exposure, particularly in public places and workplaces. However, older adults (65+) show markedly

lower levels of exposure in all environments. These findings highlight critical areas for targeted public health interventions, especially in reducing exposure for young and middle-aged women in public places and workplaces.

Figure 3a represents the entire population's passive smoking exposure patterns across regions. Public places in Makkah Al-Mokarramah exhibit the highest exposure levels, with more than 600,000 people affected. Other areas, such as Al-Riyadh and Al-Madinah Al-Monawarah, also report high exposure in public places, while smaller regions, like Northern Borders and Najran, show near-zero exposure. Homes are identified as the second-highest area of exposure, with Makkah and Al-Riyadh leading, followed by Al-Madinah. Smaller regions contribute minimally to home exposure. Workplaces show moderate exposure, particularly in Makkah, Al-Riyadh, and Al-Madinah, while other areas report significantly lower numbers. Coffee shops and cars have the lowest levels of passive smoking exposure across all regions, especially in smaller areas. However, Makkah and Al-Riyadh still report relatively higher exposure levels in these settings than in other areas. In general, Makkah Al-Mokarramah consistently exhibits the highest levels of passive smoking exposure in all environments, while smaller regions experience significantly lower levels. Public places and homes emerge as the most significant contributors to passive smoking exposure, warranting targeted interventions in these areas.

Figure 3b illustrates passive smoking exposure among the male population across various regions. Public places emerge as the most significant contributor to exposure, with a pronounced peak in Makkah Al-Mokarramah. Homes represent the second-highest source of exposure, particularly in Makkah, Al-Riyadh, and Al-Madinah. Workplaces,

coffee shops, and cars show much lower levels of exposure overall, with only modest peaks in the more prominent regions such as Makkah and Al-Riyadh. Smaller regions exhibit minimal exposure in all environments. In summary, the graph highlights that Makkah Al-Mokarramah and Al-Riyadh experience the highest levels of passive smoking exposure, particularly in public places and homes. Public places consistently stand out as the largest source of exposure, emphasizing the need for targeted interventions in these environments to reduce health risks among the male population.

Figure 3c illustrates passive smoking exposure among the female population across different regions. Public places in Makkah Al-Mokarramah exhibit the highest exposure levels, with more than 600,000 people affected. Other areas, such as Al-Riyadh and Al-Madinah Al-Monawarah, also report significant exposure in public places, whereas smaller regions, like Northern Borders and Najran, show near-zero exposure. Homes represent the second-highest source of exposure, with Makkah and Al-Riyadh leading, followed by Al-Madinah. Smaller regions contribute minimally to home exposure. Workplaces show moderate exposure levels, particularly in Makkah, Al-Riyadh, and Al-Madinah, while other areas report significantly lower numbers. Coffee shops and cars consistently have the lowest exposure levels, especially in smaller regions. However, Makkah and Al-Riyadh still report relatively higher exposure levels in these categories than in other areas. In conclusion, Makkah Al-Mokarramah stands out as the region with the highest levels of passive smoking exposure across all environments, while smaller areas show significantly lower exposure. Public places and homes are identified as the most critical areas of concern, highlighting the need for targeted public health efforts to reduce exposure in these environments among females.

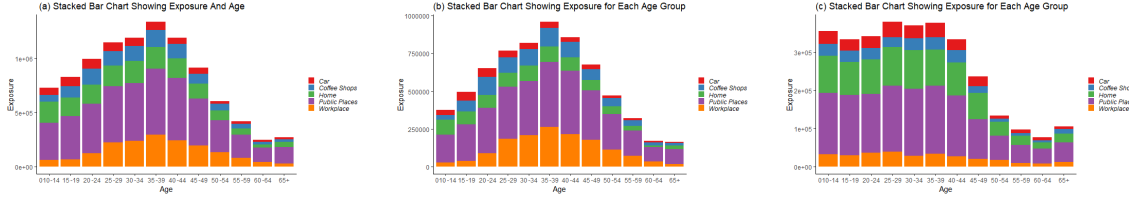


Figure 3.3: A bar chart illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and age group

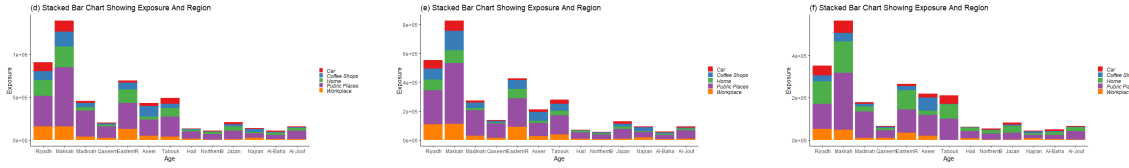


Figure 3.4: A bar chart illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and administrative area

The stacked bar charts from Figures 4a to 5f provide a comprehensive analysis of exposure patterns across age groups and regions, categorized by five places: cars, coffee shops, homes, public places, and Workplaces. Across age groups, the data reveal a consistent trend: exposure levels increase with age, peak in the 30-44 age range, and then decline in older age groups (45+). Public places consistently dominate exposure across all demographics, contributing the largest share of cumulative exposure levels. Workplace exposure plays a significant secondary role, particularly in working-age groups (20-44), reflecting the impact of professional activities during these years. Home exposure remains steady across all age groups, with slightly higher contributions from younger individuals, likely due to family living patterns, and to middle-aged individuals. Coffee shops contribute modestly to exposure, particularly in younger and middle-aged groups, while cars consistently exhibit the lowest exposure levels across all age groups. Regionally, the charts highlight significant dis-

parities in exposure levels. Urban centers such as Riyadh and Makkah exhibit the highest levels of exposures driven primarily by public places and Workplaces, which reflect their high population densities, urbanization, and vibrant social and professional activities. Moderate exposure regions, including Madinah, Qaseem, and the Eastern Region, show reduced exposure levels compared to Riyadh and Makkah but still maintain significant activity levels, with Public Places as the dominant contributor and smaller contributions from home and Workplace. Low-exposure regions, such as Hail, Jazan, Al-Baha, and Najran, demonstrate significantly lower exposure levels, driven primarily by Public Places, with minimal contributions from Workplace, Home, and Coffee Shops. These patterns are likely due to lower population densities and reduced professional and social activities in these areas. Across both age groups and regions, Public Places emerge as the most significant driver of exposure, followed by the Workplace in urban centers and working-age groups. Home contributes moderately, while coffee shops and cars play more minor roles. These findings highlight the importance of targeted strategies to manage exposure risks effectively. Interventions should prioritize reducing exposure in Public Places, particularly in high-risk urban regions, through improved ventilation, occupancy control, and other measures. For workplaces, strategies should focus on working-age groups, such as enhancing indoor air quality and promoting flexible work arrangements. Urban centers require intensive management to address high exposure levels at the regional level, while smaller regions would benefit from equitable distribution of resources to mitigate risks. These insights provide a foundation for informed decision-making to reduce exposure risks in different demographics and regions.

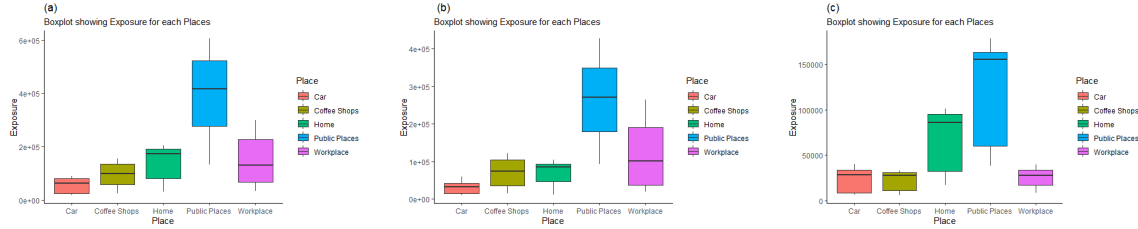


Figure 3.5: A box plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and age group

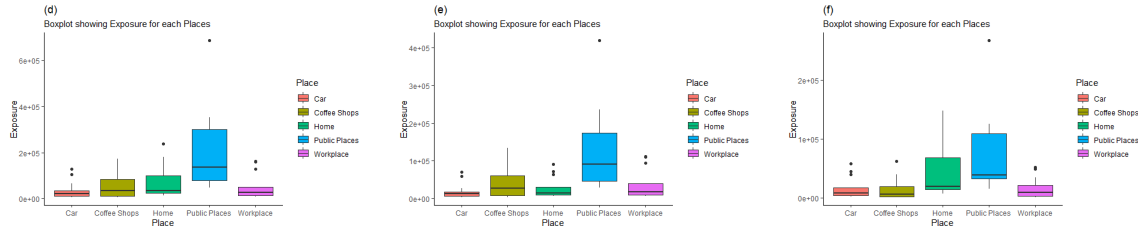


Figure 3.6: A box plot illustrating the distribution of the population aged 10 years and older exposed to passive smoking, categorized by place of exposure and administrative area

The series of boxplots in Figures 6a to 7f collectively highlight the distribution of exposure levels across five categorized locations: cars, coffee shops, homes, public Places, and Workplaces. These visualizations provide detailed insights into the variability, central tendencies, and outliers associated with exposure levels in each environment. In all figures, cars consistently exhibit the lowest exposure levels, characterized by narrow interquartile ranges (IQRs) and medians close to zero, reflecting minimal variability and highly controlled conditions. Although occasional outliers indicate rare instances of elevated exposure, cars remain the safest environment in terms of exposure. Similarly, coffee shops demonstrate slightly higher exposure levels than car shops, with broader IQRs and a few outliers, indicating moderate variability. These semi-enclosed environments generally maintain low exposure levels compared

to more dynamic settings. Homes exhibit moderate exposure levels across all figures, with wider IQRs and higher medians than those observed for cars and coffee shops. The increased variability is likely due to differences in household size, visitor interactions, and indoor activities. Occasional outliers suggest sporadic cases of elevated exposure; however, the overall exposure levels in Homes remain concentrated and controlled. In contrast, Public Places consistently emerge as the category with the highest exposure levels, displaying the largest IQRs, the highest medians, and frequent outliers. This variability highlights public places' dynamic and unpredictable nature, where diverse activities and interactions lead to frequent high-exposure events. Public places consistently stand out as the most variable and high-risk environment in all figures. Workplaces exhibit moderate exposure levels, falling between Public Places and the more controlled environments of Cars, Coffee Shops, and Homes. While the IQRs and medians for Workplace exposure are narrower and lower than those of Public Places, they still reflect significant exposure levels, likely due to professional interactions. Occasional outliers indicate sporadic elevated exposure, though these are less extreme and less frequent than in public places. Workplace exposure appears more regulated and consistent than the variability observed in public environments. While the general trends remain consistent across all figures, the differences in exposure scales and the presence of outliers emphasize variability across locations. Figures (a), (b), and (c) focus on more minor scales, capturing subtle differences in exposure, while figures (d), (e), and (f) extend the range to highlight more prominent outliers. These visualizations consistently demonstrate that Public Places pose the highest exposure risks due to their openness, high variability, and elevated medians. In contrast,

cars and coffee shops provide the safest environments with consistently low exposure. Homes and Workplaces occupy intermediate positions, varyingly influenced by household dynamics and workplace structures. The presence of outliers in categories such as Public Places and Homes underscores the need for targeted interventions to mitigate risks in these settings. These findings are essential for understanding exposure patterns and informing strategies to manage exposure risks, particularly in high-risk environments such as Public Places and Workplaces.

Table 3.1: Summary and Descriptive Statistics of the Entire Population by Age Group

STATISTIC	Car	Coffee Shops	Home	Public Places	Workplaces
Nbrval	12	12	12	12	12
Min.val	16311	23858	29719	131873	32270
Q1	24127.75	57589.25	79917	275535.2	66361
Sum	674113	1119093	1668925	4651871	1749738
Median	63143	9784	173027.5	416096.5	130004.5
Mean	56176.08	93257.75	139077.1	387655.9	145811.5
Q3	80207	133738.5	191849.5	522422	227190
Max.val	89788	155859	205061	606014	298161
Skewness	-0.3210183	-0.1448005	-0.5459293	-0.2659627	0.2245061
Kurtosis	-1.738436	-1.710605	-1.552718	-1.472133	-1.623495
Std.dev	28725.74	48196.58	65404.72	162997.1	90636.22
SE mean	8292.407	13913.15	18880.72	47053.21	26164.42
Variance	825168189	2322910257	4277777032	26568055252	8214923530
CV	0.5113518	0.5168104	0.4702767	0.4204685	
0.6215985 Range	73477	132001	175342	474141	265891
IQR	56079.25	76149.25	111932.5	246886.8	160829

Table 3.2: Summary and Descriptive Statistics of the Male Population by Age Group

STATISTIC	Car	Coffee Shops	Home	Public Places	Workplaces
Nbrval	12	12	12	12	12
Min.val	9690	14349	12616	93441	19548
Q1	15649.75	35891.5	47376	179154.8	37651.25
Sum	384197	853794	845464	3192989	1449866
Median	32447	73732.5	84656	271325.5	102134.5
Mean	32016.42	71149.5	70455.33	266082.4	120822.2
Q3	41768.5	104601	93070.5	348895.8	190389
Max.val	59053	122799	104149	427680	264651
Skewness	0.08371616	-0.1205947	-0.5805154	-0.1231312	0.2372515
Kurtosis	-1.453741	-1.705229	-1.360218	-1.483402	-1.640516
Std.dev	17112.12	39540.89	32170.87	115194.2	85653.37
SE mean	4939.844	11414.47	9286.93	33253.71	24726
Variance	292824749	1563481988	1034964801	13269710735	7336499391
CV	0.5344796	0.5557438	0.4566137	0.4329269	0.708921
Range	49363	108450	91533	334239	245103
IQR	26118.75	68709.5	45694.5	169741	152737.8

Table 3.3: Summary and Descriptive Statistics of the Female Population by Age Group

STATISTIC	Car	Coffee Shops	Home	Public Places	Workplaces
Nbrval	12	12	12	12	12
Min.val	6621	5562	17103	38432	8692
Q1	8494	11166.75	32541	60276.25	16945.25
Sum	289916	265299	823461	1458882	299872
Median	28471.5	27840	85991	155327.5	27897.5
Mean	24159.67	22108.25	68621.75	121573.5	24989.33
Q3	33760.25	31014	94732	162851.5	33312
Max.val	40081	33060	100912	178334	39563
Skewness	-0.3665944	-0.401127	-0.5030653	-0.4112813	-0.2270693
Kurtosis	-1.697889	-1.764158	-1.684598	-1.78763	-1.56235
Std.dev	12848.23	10993.87	33651.11	56475.32	10614.11
SE mean	3708.965	3173.658	9714.239	16303.02	3064.03
Variance	165077076	120865269	1132397352	3189462082	112659373
CV	0.531805	0.4972747	0.490385	0.4645365	0.4247457
Range	33460	27498	83809	139902	30871
IQR	25266.25	19847.25	62191	102575.2	16366.75

Table 3.4: Summary and Descriptive Statistics of the Entire Population by Region

STATISTIC	Car	Coffee Shops	Home	Public Places	Workplaces
Nbrval	13	13	13	13	13
Min.val	4866	5714	13373	47447	8873
Q1	10496	12077	24036	80051	12499
Sum	465839	697544	926587	2642892	700038
Median	22301	35626	34610	135631	26959
Mean	35833.77	53657.23	71275.92	203299.4	53849.08
Q3	34451	83099	99228	299844	50199
Max.val	128579	174113	239233	687263	161959
Skewness	1.285114	0.934442	1.107497	1.392477	1.039928
Kurtosis	0.1355987	-0.4267709	-0.342069	1.283822	-0.7079611
Std.dev	39597.55	52885.65	73789.05	179287.1	56993.64
SE mean	10982.38	14667.84	20465.4	49725.29	15807.19
Variance	1567965966	2796892004	5444823166	32143851835	3248275115
CV	1.105034	0.9856202	1.035259	0.8818869	1.058396
Range	123713	168399	225860	639816	153086
IQR	23955	71022	75192	219793	37700

Table 3.5: Summary and Descriptive Statistics of the Male Population by Region

STATISTIC	Car	Coffee Shops	Home	Public Places	Workplaces
Nbrval	13	13	13	13	13
Min.val	2426	3228	5688	28805	5437
Q1	6038	8244	9346	45510	9418
Sum	248550	503405	354587	1621473	487796
Median	12517	26246	14207	90716	17917
Mean	19119.23	38723.46	27275.92	124728.7	37522.77
Q3	17128	61075	30478	174273	38927
Max.val	70368	134324	90816	419970	111348
Skewness	1.405703	1.092831	1.127499	1.371167	0.9736535
Kurtosis	0.4981782	0.3671473	-0.4335206	1.101491	-0.8537758
Std.dev	21598.87	38029.56	28829.66	110769.1	39826.94
SE mean	5965.487	10547.5	7995.91	30721.83	11046.01
Variance	462631433	1446247634	831149556	12269803783	1586185005
CV	1.124986	0.9820807	1.056964	0.8880807	1.061407
Range	67942	131096	85128	390865	105911
IQR	11090	52831	21132	128763	29509

Table 3.6: Summary and Descriptive Statistics of the Female Population by Region

STATISTIC	Car	Coffee Shops	Home	Public Places	Workplaces
Nbrval	13	13	13	13	13
Min.val	2440	1063	7685	15149	943
Q1	4841	2503	14805	32494	3440
Sum	217289	194139	572000	1021419	212242
Median	7997	6154	19858	38767	9042
Mean	16714.54	14933.77	44000	78570.69	16326.31
Q3	17529	19835	68750	109970	21626
Max.val	58211	63010	148417	267593	52679
Skewness	1.148377	1.408602	1.094426	1.35488	1.039909
Kurtosis	-0.3598066	0.8411931	-0.2975101	1.275246	-0.5739225
Std.dev	18576.67	18536.43	45184.18	70003.61	17930.77
SE mean	5152.24	5141.082	12531.84	19415.51	4973.101
Variance	345092493	343599388	2041610017	4900505231	32152521
CV	1.111408	1.241243	1.026913	0.8909634	1.098275
Range	55771	619447	140732	252444	51736
IQR	12688	17332	53945	77476	18186

The six tables collectively provide a detailed breakdown of descriptive statistics in various demographic and regional segments, emphasizing central tendencies, variability, and distribution characteristics. Table 1 highlights the entire population in five categories: cars, coffee shops, homes, public places, and workplaces. It demonstrates that public places exhibit the most tremendous variability, as indicated by the highest standard deviation (162,997.1) and range (474,411), while cars show relatively low variability with an IQR of 56,079.25. Skewness and kurtosis reveal mostly symmetric and flat distributions across categories, while the coefficient of variation (CV) normalizes dispersion, showing public places as having the lowest relative variability (CV = 0.4205). These findings provide a general overview of population-wide trends in the five categories. Tables 2 and 3 delve deeper into male and female populations, offering gender-specific insights. Among males, public places report the highest mean (266,082.4) and variability (Std.dev = 115,194.2), while cars show the lowest mean (32,016.42) and variability (Std.dev = 17,112.12). Skewness values indicate that cars are nearly symmetric (0.083), while Workplaces are slightly right-skewed (0.237). For females, home has the highest mean (68,621.75), while coffee shops show the lowest (22,108.25). Variability is most prominent in public places (Std.dev = 56,475.32), and the distributions show slight left skewness except for Workplaces, which approach symmetry (-0.227). Both tables highlight gender-specific differences, with public places consistently demonstrating the highest variability in both groups, reflecting a broader spread and distribution.

The regional analysis in Tables 4 through 6 provides additional context by focusing on

the entire population, male population, and female population segmented by region. Table 4 shows that public places have the highest mean (203,299.4) and variability (Std.dev = 179,287.1) in all areas, while cars exhibit the lowest mean (35,837.77) and variability (IQR = 23,955). Table 5 identifies similar patterns among males, with public places showing the highest variability (Std.dev = 110,769.1) and skewness (1.3712), while cars remain the least variable (IQR = 11,090). Finally, Table 6 for females reveals that public places also dominate in mean values (78,570.69) and standard deviation (70,003.61), while cars have the smallest range (55,771). Skewness and kurtosis values indicate flatter and positively skewed distributions across categories, particularly for females in coffee shops and cars. These insights collectively showcase the consistency of trends in gender and regional populations, highlighting significant patterns in data distribution and variability for each category.

3.4 Estimated Results (Fitting the Multinomial Model)

Table 3.7: Fitting the Multinomial Model to the Entire Population Stratified by Region

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
(Intercept)	10.447293	-	0.005388	1939.120	$< 2e-16$
Coffee Shops	1.281429	3.60	0.006090	210.422	$< 2e-16$
Home	0.004605	1.00	0.007611	0.605	0.545156
Public Places	1.692618	5.43	0.005862	288.720	$< 2e-16$
Workplace	0.376457	1.46	0.006996	53.808	$< 2e-16$
Al-Baha	-0.867460	0.420	0.009906	-87.566	$< 2e-16$
Eastern Region	-0.434907	0.647	0.008595	-50.602	$< 2e-16$
Hail	-1.957266	0.141	0.015315	-127.805	$< 2e-16$
Jazan	-0.245555	0.782	0.008132	-30.195	$< 2e-16$
Al-Jouf	-1.105486	0.331	0.010803	-102.330	$< 2e-16$
Al-Madinah	-0.397975	0.672	0.008500	-46.823	$< 2e-16$
Makkah	1.317005	3.73	0.006067	217.090	$< 2e-16$
Najran	-1.418835	0.242	0.012205	-116.246	$< 2e-16$
Northern Borders	-1.293206	0.274	0.011611	-111.379	$< 2e-16$
Al-Qaseem	-1.188544	0.305	0.011149	-106.605	$< 2e-16$
Riyad	1.110032	3.030	0.006212	178.683	$< 2e-16$
Tabouk	0.662928	1.940	0.006632	99.957	$< 2e-16$
CoffeeShops:Al-Baha	-1.462204	0.232	0.013748	-106.361	$< 2e-16$
Home:Al-Baha	0.492163	1.640	0.013003	37.850	$< 2e-16$
PublicPlaces:Al-Baha	-0.505082	0.603	0.011160	-45.257	$< 2e-16$
Workplace:Al-Baha	-0.544397	0.580	0.014135	-38.514	$< 2e-16$
CoffeeShops:Eastern Region	0.033972	1.030	0.009693	3.505	0.000457
Home:Eastern Region	1.930641	6.890	0.010452	184.722	$< 2e-16$
PublicPlaces:Eastern Region	0.921589	2.510	0.009083	101.468	$< 2e-16$
Workplace:Eastern Region	1.370590	3.940	0.010080	135.973	$< 2e-16$
CoffeeShops:Hail	-1.120782	0.326	0.020435	-54.845	$< 2e-16$
Home:Hail	1.592676	4.920	0.017465	91.192	$< 2e-16$
Public Places:Hail	1.107774	3.030	0.015886	69.732	$< 2e-16$
Workplace:Hail	0.960770	2.610	0.017562	54.708	$< 2e-16$

Table 3.8: Continuation of Table 13

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
CoffeeShops:Jazan	-1.378373	0.252	0.010727	-128.492	$< 2e-16$
Home:Jazan	0.572509	1.770	0.010764	53.188	$< 2e-16$
PublicPlaces:Jazan	-0.385208	0.680	0.009028	-42.666	$< 2e-16$
Workplace:Jazan	-1.205141	0.300	0.013078	-92.150	$< 2e-16$
CoffeeShops:Al-Jouf	-1.484714	0.227	0.015239	-97.432	$< 2e-16$
Home:Al-Jouf	0.939579	2.560	0.013406	70.089	$< 2e-16$
Public Places:Al-Jouf	0.462178	1.590	0.011498	40.195	$< 2e-16$
Workplace:Al-Jouf	-0.284860	0.752	0.014719	-19.354	$< 2e-16$
Coffee Shops:Al-Madinah	-0.495863	0.609	0.010000	-49.586	$< 2e-16$
Home:Al-Madinah	0.494650	1.640	0.011286	43.830	$< 2e-16$
Public Places:Al-Madinah	0.869081	2.380	0.008995	96.613	$< 2e-16$
Workplace:Al-Madinah	0.221458	1.250	0.010767	20.568	$< 2e-16$
Coffee Shops:Makkah	-0.978268	0.376	0.007114	-137.517	$< 2e-16$
Home:Makkah	0.616290	1.850	0.008359	73.725	$< 2e-16$
Public Places:Makkah	-0.016445	0.984	0.006603	-2.490	0.012759
Workplace:Makkah	-0.151683	0.859	0.007933	-19.120	$< 2e-16$
Coffee Shops:Najran	0.170943	1.19	0.013605	12.565	$< 2e-16$
Home:Najran	0.467930	1.600	0.015895	29.439	$< 2e-16$
PublicPlaces:Najran	0.186749	1.210	0.013139	14.213	$< 2e-16$
Workplace:Najran	0.797157	2.220	0.014352	55.542	$< 2e-16$
CoffeeShops:Northern Borders	-1.554375	0.211	0.016789	-92.585	$< 2e-16$
Home:Northern Borders	0.973484	2.650	0.014265	68.244	$< 2e-16$
PublicPlaces:Northern Borders	0.174298	1.190	0.012510	13.932	$< 2e-16$
Workplace:Northern Borders	-0.439776	0.644	0.016353	-26.892	$< 2e-16$
CoffeeShops:Al-Qaseem	-0.822238	0.439	0.013876	-59.254	$< 2e-16$
Home:l-Qaseem	0.498685	1.650	0.014519	34.348	$< 2e-16$
PublicPlaces:Al-Qaseem	0.866326	2.380	0.011705	74.011	$< 2e-16$
Workplace:Al-Qaseem	0.535633	1.710	0.013509	39.650	$< 2e-16$
CoffeeShops:Riyad	-1.289284	0.275	0.007503	-171.840	$< 2e-16$
Home:Riyad	0.538438	1.710	0.008547	63.001	$< 2e-16$
PublicPlaces:Riyad	-0.473634	0.623	0.006838	-69.263	$< 2e-16$
Workplace:Riyad	0.061316	1.060	0.008043	7.624	$2.47e-14$
CoffeeShops:Tabouk	-1.556550	0.211	0.008470	-183.773	$< 2e-16$
Home:Tabouk	0.390349	1.480	0.009108	42.858	$< 2e-16$
PublicPlaces:Tabouk	-0.442241	0.643	0.007322	-60.399	$< 2e-16$
Workplace:Tabouk	-0.893299	0.409	0.009433	-94.696	$< 2e-16$

Table 3.9: **Fitting the Multinomial Model to the Male Population Stratified by Region**

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
(Intercept)	9.736370	-	0.007687	1266.552	$< 2e-16$
CoffeeShops	1.283488	3.610	0.008687	147.744	$< 2e-16$
Home	-0.137236	0.872	0.011264	-12.183	$< 2e-16$
PublicPlaces	1.679119	5.360	0.008374	200.524	$< 2e-16$
Workplace	0.523848	1.690	0.009700	54.004	$< 2e-16$
Al-Baha	-0.882562	0.414	0.014210	-62.107	$< 2e-16$
Eastern Region	-0.301527	0.740	0.011789	-25.57	$< 2e-16$
Hail	-1.942371	0.143	0.021709	-89.472	$< 2e-16$
Jazan	0.012100	1.010	0.010839	1.116	0.264
Al-Jouf	-0.947014	0.388	0.014541	-65.127	$< 2e-16$
Al-Madinah	-0.111076	0.895	0.011186	-9.930	$< 2e-16$
Makkah	1.425124	4.160	0.008562	166.450	$< 2e-16$
Najran	-1.275112	0.279	0.016450	-77.515	$< 2e-16$
Northern Borders	-1.477170	0.228	0.017831	-82.841	$< 2e-16$
Al-Qaseem	-1.030542	0.357	0.014990	-68.747	$< 2e-16$
Riyad	1.252206	3.500	0.008717	143.649	$< 2e-16$
Tabouk	0.454475	1.580	0.009829	46.239	$< 2e-16$
CoffeeShops:Al-Baha	-1.173240	0.309	0.018607	-63.053	$< 2e-16$
Home:Al-Baha	0.426131	1.530	0.019409	21.955	$< 2e-16$
PublicPlaces:Al-Baha	-0.264623	0.767	0.015738	-16.815	$< 2e-16$
Workplace:Al-Baha	-0.548137	0.578	0.019577	-27.998	$< 2e-16$
CoffeeShops:Eastern Region	0.336741	1.400	0.013083	25.739	$< 2e-16$
Home:Eastern Region	1.767280	5.850	0.014914	118.498	$< 2e-16$
PublicPlaces:Eastern Region	1.064642	2.900	0.012456	85.473	$< 2e-16$
Workplace:Eastern Region	1.486480	4.420	0.013590	109.381	$< 2e-16$
CoffeeShops:Hail	-0.997869	0.369	0.028239	-35.336	$< 2e-16$
Home:Hail	1.473559	4.360	0.025444	57.913	$< 2e-16$
PublicPlaces:Hail	1.252569	3.500	0.022456	55.778	$< 2e-16$
Workplace:Hail	0.856141	2.350	0.024698	34.664	$< 2e-16$
CoffeeShops:Jazan	-1.216974	0.296	0.013729	-88.644	$< 2e-16$
Home:Jazan	0.128675	1.140	0.015625	8.235	$< 2e-16$
PublicPlaces:Jazan	-0.313218	0.731	0.011975	-26.156	$< 2e-16$
Workplace:Jazan	-1.121939	0.326	0.016083	-69.760	$< 2e-16$

Table 3.10: Continuation of Table 15

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
CoffeeShops:Al-Jouf	-1.055603	0.348	0.018685	-56.496	$< 2e-16$
Home:Al-Jouf	0.585252	1.800	0.019405	30.160	$< 2e-16$
PublicPlaces:Al-Jouf	0.527171	1.690	0.015467	34.083	$< 2e-16$
Workplace:Al-Jouf	-0.201689	0.817	0.018890	-10.677	$< 2e-16$
CoffeeShops:Al-Madinah	-0.332630	0.717	0.012924	-25.738	$< 2e-16$
Home:Al-Madinah	0.073432	1.080	0.016227	4.525	$6.03e-06$
PublicPlaces:Al-Madinah	0.763965	2.150	0.011912	64.135	$< 2e-16$
Workplace:Al-Madinah	0.218960	1.240	0.013840	15.821	$< 2e-16$
CoffeeShops:Makkah	-0.636972	0.529	0.009855	-64.634	$< 2e-16$
Home:Makkah	0.392333	1.480	0.012333	31.811	$< 2e-16$
PublicPlaces:Makkah	0.106611	1.110	0.009312	11.449	$< 2e-16$
Workplace:Makkah	-0.064926	0.937	0.010830	-5.995	$2.03e-09$
CoffeeShops:Najran	0.430523	1.540	0.018030	23.878	$< 2e-16$
Home:Najran	0.322092	1.380	0.022676	14.204	$< 2e-16$
PublicPlaces:Najran	0.442489	1.560	0.017521	25.255	$< 2e-16$
PlacesWorkplace:Najran	0.808400	2.240	0.019011	42.523	$< 2e-16$
CoffeeShops:Northern Borders	-1.089073	0.337	0.023398	-46.546	$< 2e-16$
Home:Northern Borders	1.041600	2.830	0.022148	47.029	$< 2e-16$
PublicPlaces:Northern Borders	0.490962	1.630	0.018935	25.929	$< 2e-16$
Workplace:Northern	-0.182064	0.834	0.023171	-7.858	$3.92e-15$
CoffeeShops:Al-Qaseem	-0.617282	0.539	0.018060	-34.180	$< 2e-16$
Home:Al-Qaseem	0.296154	1.340	0.020827	14.219	$< 2e-16$
Publi Places:Al-Qaseem	1.110089	3.030	0.015682	70.789	$< 2e-16$
Workplace:Al-Qaseem	0.455904	1.580	0.017941	25.411	$< 2e-16$
CoffeeShops:Riyad	-1.030973	0.357	0.010271	-100.381	$< 2e-16$
Home:RegionRiyad	0.337295	1.400	0.012554	26.867	$< 2e-16$
PublicPlaces:Riyad	-0.295591	0.744	0.009552	-30.944	$< 2e-16$
Workplace:Riyad	0.089246	1.090	0.010961	8.142	$3.88e-16$
Coffee Shops:Tabouk	-0.673391	0.510	0.011548	-58.311	$< 2e-16$
Home:Tabouk	0.271152	1.310	0.014043	19.309	$< 2e-16$
Public Places:Tabouk	-0.067830	0.934	0.010729	-6.322	$2.58e-10$
Workplace:Tabouk	-0.145249	0.865	0.012542	-11.581	$< 2e-16$

Table 3.11: **Fitting the Multinomial Model to the Female Population Stratified by Region**

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
(Intercept)	9.771612	-	0.007553	1293.733	$< 2e-16$
Coffee Shops	1.279437	3.590	0.008539	149.830	$< 2e-16$
Home	0.124750	1.130	0.010364	12.037	$< 2e-16$
Public Places	1.705479	5.500	0.008211	207.718	$< 2e-16$
Workplace	0.210040	1.230	0.010163	20.667	$< 2e-16$
Al-Baha	-0.853096	0.426	0.013818	-61.738	$< 2e-16$
Eastern Region	-0.583108	0.558	0.012620	-46.206	$< 2e-16$
Hail	-1.971859	0.139	0.021608	-91.258	$< 2e-16$
Jazan	-0.579232	0.560	0.012604	-45.956	$< 2e-16$
Al-Jouf	-1.286735	0.276	0.016236	-79.251	$< 2e-16$
Al-Madinah	-0.784790	0.456	0.013494	-58.157	$< 2e-16$
Makkah	1.200218	3.320	0.008616	139.309	$< 2e-16$
Najran	-1.580426	0.206	0.018279	-86.460	$< 2e-16$
Northern Borders	-1.142877	0.319	0.015360	-74.405	$< 2e-16$
Al-Qaseem	-1.369156	0.254	0.016774	-81.624	$< 2e-16$
Riyad	0.950465	2.590	0.008894	106.867	$< 2e-16$
Tabouk	0.829836	2.290	0.009051	91.680	$< 2e-16$
CoffeeShops:Al-Baha	-1.840929	0.159	0.021013	-87.609	$< 2e-16$
Home:Al-Baha	0.534075	1.710	0.017623	30.305	$< 2e-16$
PublicPlaces:Al-Baha	-0.790823	0.453	0.015967	-49.529	$< 2e-16$
Workplace:Al-Baha	-0.534402	0.586	0.020551	-26.003	$< 2e-16$
CoffeeShops:Eastern Region	-0.572737	0.564	0.015018	-38.137	$< 2e-16$
Home:Eastern Region	2.100547	8.170	0.014854	141.408	$< 2e-16$
PublicPlaces:Eastern Region	0.713980	2.040	0.013368	53.408	$< 2e-16$
Workplace:Eastern Region	1.050345	2.860	0.015313	68.593	$< 2e-16$
CoffeeShops:Hail	-1.260760	2.283	0.029749	-42.380	$< 2e-16$
Home:Hail	1.678217	5.360	0.024182	69.398	$< 2e-16$
PublicPlaces:Hail	0.944670	2.570	0.022499	41.987	$< 2e-16$
Workplace:Hail	1.082777	2.950	0.025012	43.289	$< 2e-16$
CoffeeShops:Jazan	-1.746959	0.174	0.018364	-95.131	$< 2e-16$
Home:Jazan	1.025031	2.790	0.015539	65.965	$< 2e-16$
PublicPlaces:Jazan	-0.509048	0.601	0.014142	-35.995	$< 2e-16$
Workplace:Jazan	-1.640675	0.194	0.025116	-65.324	$< 2e-16$

Table 3.12: Continuation of Table 17

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
CoffeeShops:Al-Jouf	-2.795463	0.061	0.034932	-80.027	$< 2e-16$
Home:Al-Jouf	1.244933	3.470	0.019144	65.031	$< 2e-16$
PublicPlaces:Al-Jouf	0.374969	1.450	0.017314	21.657	$< 2e-16$
Workplace:Al-Jouf	-0.551689	0.576	0.024506	-22.512	$< 2e-16$
CoffeeShops:Al-Madinah	-0.909915	0.403	0.016864	-53.957	$< 2e-16$
Home:Al-Madinah	0.970731	2.640	0.016561	58.616	$< 2e-16$
PublicPlaces:Al-Madinah	1.048326	2.850	0.014157	74.049	$< 2e-16$
Workplace:Al-Madinah	0.037976	1.040	0.018053	2.104	0.0354
CoffeeShops:Makkah	-1.659921	0.190	0.010735	-154.634	$< 2e-16$
Home:Makkah	0.811201	2.250	0.011460	70.788	$< 2e-16$
PublicPlaces:Makkah	-0.180086	0.835	0.009398	-19.161	$< 2e-16$
Workplace:Makkah	-0.369357	0.691	0.011858	-31.148	$2.03e-09$
CoffeeShops:Najran	0.430523	0.723	0.018030	23.878	$< 2e-16$
Home:Najran	0.322092	1.880	0.022676	14.204	$< 2e-16$
PublicPlaces:Najran	0.442489	0.763	0.017521	25.255	$< 2e-16$
Workplace:Najran	0.808400	2.030	0.019011	42.523	$< 2e-16$
CoffeeShops:Northern Borders	-1.089073	0.125	0.023398	-46.546	$< 2e-16$
Home:Northern Borders	1.041600	2.460	0.022148	47.029	$< 2e-16$
PublicPlaces:Northern Borders	0.490962	0.888	0.018935	25.929	$< 2e-16$
Workplace:Northern Borders	-0.182064	0.498	0.023171	-7.858	$3.92e-15$
CoffeeShops:Al-Qaseem	-1.193510	0.303	0.022429	-53.212	$< 2e-16$
Home:Al-Qaseem	0.711139	2.040	0.020711	34.336	$< 2e-16$
PublicPlaces:Al-Qaseem	0.421598	1.520	0.017846	23.625	$< 2e-16$
Workplace:Al-Qaseem	0.602435	1.83	0.020668	29.149	$< 2e-16$
CoffeeShops:Riyad	-1.778700	0.169	0.011459	-155.226	$< 2e-16$
Home:Riyad	0.739664	2.100	0.011779	62.794	$< 2e-16$
PublicPlaces:Riyad	-0.752384	0.471	0.009898	-76.015	$< 2e-16$
Workplace:Riyad	-0.060145	0.942	0.012014	-5.006	$5.55e-07$
CoffeeShops:Tabouk	-4.439565	0.012	0.026159	-169.713	$< 2e-16$
Home:Tabouk	0.412034	1.510	0.012117	34.003	$< 2e-16$
Public Places:Tabouk	-0.795162	0.452	0.010115	-78.615	$< 2e-16$
Workplace:Tabouk	3.962421	0.019	0.034476	-114.932	$< 2e-16$

Table 3.13: **Fitting the Multinomial Model to the Entire Population Stratified by Age Group**

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
(Intercept)	11.120357	-	0.003848	2889.838	< 2e-16
Coffee Shops	-0.086888	0.917	0.005564	-15.616	< 2e-16
Home	1.063353	2.900	0.004463	238.245	< 2e-16
Public Places	1.626006	5.080	0.004210	386.262	< 2e-16
Workplace	-0.096479	0.908	0.005578	-17.296	< 2e-16
15-19	0.237177	1.270	0.005147	46.083	< 2e-16
20-24	0.284850	1.330	0.005094	55.923	< 2e-16
25-29	0.215521	1.240	0.005172	41.675	< 2e-16
30-34	0.104780	1.110	0.005305	19.751	< 2e-16
35-39	0.157074	1.170	0.005241	29.973	< 2e-16
40-44	-0.139242	0.870	0.005642	-24.681	< 2e-16
45-49	-0.206087	0.814	0.005745	-35.873	< 2e-16
50-54	-0.987386	0.373	0.007386	-133.681	< 2e-16
55-59	-1.166461	0.311	0.007896	-147.725	< 2e-16
60-64	-1.368390	0.255	0.008543	-160.171	< 2e-16
65+	-1.420762	0.242	0.008724	-162.848	< 2e-16
CoffeeShops:15-19	0.278264	1.320	0.007231	38.482	< 2e-16
Home:15-19	-0.374203	0.688	0.006121	-61.135	< 2e-16
Public Places:15-19	-0.082694	0.921	0.005648	-14.642	< 2e-16
Workplace:15-19	-0.133189	0.875	0.007582	-17.565	< 2e-16
CoffeeShops:20-24	0.592080	1.810	0.006987	84.741	< 2e-16
Home:20-24	-0.393040	0.675	0.006063	-64.830	< 2e-16
PublicPlaces:20-24	-0.008327	0.992	0.005574	-1.494	0.13518
Workplace:20-24	0.442380	1.560	0.007080	62.479	< 2e-16
CoffeeShops:25-29	0.526015	1.690	0.007113	73.954	< 2e-16
Home:25-29	-0.241262	0.786	0.006091	-39.607	< 2e-16
PublicPlaces:25-29	0.198207	1.220	0.005620	35.269	< 2e-16
Workplace:25-29	1.077916	2.940	0.006894	156.358	< 2e-16
CoffeeShops:30-34	0.692446	2.000	0.007181	96.422	< 2e-16
Home:30-34	-0.057427	0.944	0.006175	-9.300	< 2e-16
PublicPlaces:30-34	0.333286	1.400	0.005739	58.075	< 2e-16
Workplace:30-34	1.252045	3.500	0.006975	179.508	< 2e-16

Table 3.14: Continuation of Table 19

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
CoffeeShops:35-39	0.766164	2.150	0.007073	108.318	< 2e-16
Home:35-39	-0.150495	0.860	0.006136	-24.525	< 2e-16
PublicPlaces:35-39	0.411221	1.510	0.005659	72.665	< 2e-16
Workplace:35-39	1.424437	4.160	0.006865	207.501	< 2e-16
CoffeeShops:40-44	0.900232	2.460	0.007452	120.811	< 2e-16
Home:40-44	0.040404	1.040	0.006526	6.191	5.96e-10
PublicPlaces:40-44	0.659907	1.930	0.006039	109.271	< 2e-16
Workplace:40-44	1.512789	4.540	0.007229	209.254	< 2e-16
CoffeeShops:45-49	0.602487	1.830	0.007747	77.766	< 2e-16
Home:45-49	-0.159332	0.853	0.006744	-23.625	< 2e-16
PublicPlaces:45-49	0.434844	1.540	0.006183	70.3242	< 2e-16
Workplace:45-49	1.374680	3.950	0.007374	186.414	< 2e-16
CoffeeShops:50-54	0.986255	2.680	0.009321	105.812	< 2e-16
Home:50-54	0.189518	1.210	0.008428	22.488	< 2e-16
PublicPlaces:50-54	0.841247	2.320	0.007800	107.853	< 2e-16
Workplace:50-54	1.762477	5.830	0.008853	199.083	< 2e-16
CoffeeShops:55-59	0.843647	2.320	0.010040	84.026	< 2e-16
Home:55-59	-0.093650	0.911	0.009246	-10.129	< 2e-16
PublicPlaces:55-59	0.685694	1.990	0.008365	81.971	< 2e-16
Workplace:55-59	1.484056	4.410	0.009515	155.977	< 2e-16
CoffeeShops:60-64	0.414796	1.510	0.011448	36.233	< 2e-16
Home:60-64	-0.515777	0.597	0.010571	-48.791	< 2e-16
Public Places:60-64	0.411622	1.510	0.009137	45.050	< 2e-16
Workplace:60-64	1.048274	2.850	0.010571	99.161	< 2e-16
CoffeeShops:65+	0.568639	1.770	0.011408	49.846	< 2e-16
Home:65+	0.028822	1.030	0.010089	2.857	0.00428
PublicPlaces:65+	0.578542	1.780	0.009262	62.462	< 2e-16
Workplace:65+	0.778777	2.180	0.011109	70.103	< 2e-16

Table 3.15: **Fitting the Multinomial Model to the Male Population Stratified by Age Group**

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
(Intercept)	10.431996	-	0.005429	1921.527	< 2e-16
CoffeeShops	-0.094423	0.910	0.007866	-12.005	< 2e-16
Home	1.066962	2.910	0.006294	169.520	< 2e-16
PublicPlaces	1.689526	5.420	0.005909	285.928	< 2e-16
Workplace	-0.189361	0.827	0.008068	-23.471	< 2e-16
15-19	0.509168	1.660	0.006869	74.122	< 2e-16
20-24	0.554195	1.740	0.006812	81.351	< 2e-16
25-29	0.252947	1.290	0.007236	34.956	< 2e-16
30-34	0.183534	1.200	0.007349	24.974	< 2e-16
35-39	0.192424	1.210	0.007334	26.237	< 2e-16
40-44	-0.091351	0.913	0.007859	-11.623	< 2e-16
45-49	-0.158118	0.854	0.008000	-19.765	< 2e-16
50-54	-0.709491	0.492	0.009455	-75.040	< 2e-16
55-59	-0.995955	0.369	0.010453	-95.277	< 2e-16
60-64	-1.188511	0.305	0.011235	-105.7911	< 2e-16
65+	-1.253146	0.286	0.011518	-108.795	< 2e-16
CoffeeShops:15-19	0.358464	1.430	0.009653	37.1342	< 2e-16
Home:15-19	-0.663631	0.515	0.008316	-79.798	< 2e-16
PublicPlaces:15-19	-0.231108	0.794	0.007533	-30.679	< 2e-16
Workplace:15-19	-0.199695	0.819	0.010438	-19.132	< 2e-16
CoffeeShops:20-24	0.795851	2.220	0.009338	85.227	< 2e-16
Home:20-24	-0.704949	0.494	0.008267	-85.275	< 2e-16
PublicPlaces:20-24	-0.064342	0.938	0.007429	-8.661	< 2e-16
Workplace:20-24	0.605243	1.830	0.009654	62.694	< 2e-16
CoffeeShops:25-29	0.961636	2.620	0.009714	98.991	< 2e-16
Home:25-29	-0.347457	0.706	0.008582	-40.488	< 2e-16
PublicPlaces:25-29	0.380782	1.460	0.007790	48.879	< 2e-16
Workplace:25-29	1.626946	5.090	0.009665	168.331	< 2e-16
CoffeeShops:30-34	1.053864	2.870	0.009788	107.673	< 2e-16
Home:30-34	-0.128915	0.879	0.008588	-15.012 0	< 2e-16
PublicPlaces:30-34	0.478991	1.610	0.007890	60.7085	< 2e-16
Workplace:30-34	1.827007	6.220	0.009716	188.047	< 2e-16

Table 3.16: Continuation of Table 21

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
CoffeeShops:35-39	1.188307	3.280	0.009712	122.353	< 2e-16
Home:35-39	-0.150064	0.861	0.008582	-17.486	< 2e-16
PublicPlaces:35-39	0.652184	1.920	0.007847	83.117	< 2e-16
Workplace:35-39	2.051108	7.780	0.009653	212.477	< 2e-16
CoffeeShops:40-44	1.262826	3.540	0.010208	123.712	< 2e-16
Home:40-44	0.013455	1.010	0.009103	1.478	0.139
PublicPlaces:40-44	0.915520	2.500	0.008342	109.743	< 2e-16
Workplace:40-44	2.126630	8.390	0.010101	210.527	< 2e-16
CoffeeShops:45-49	1.031730	2.810	0.010484	98.411	< 2e-16
Home:45-49	-0.225710	0.798	0.009435	-23.922	< 2e-16
PublicPlaces:45-49	0.737054	2.090	0.008514	86.570	< 2e-16
Workplace:45-49	1.999298	7.380	0.010260	194.864	< 2e-16
CoffeeShops:50-54	1.256473	3.510	0.011855	105.990	< 2e-16
Home:50-54	0.079558	1.080	0.010889	7.306	2.75e-13
PublicPlaces:50-54	0.947312	2.580	0.009956	95.149	< 2e-16
Workplace:50-54	2.117472	8.310	0.011564	183.109	< 2e-16
CoffeeShops:55-59	1.192211	3.290	0.012972	91.904	< 2e-16
Home:55-59	-0.131639	0.877	0.012277	-10.722	< 2e-16
PublicPlaces:55-59	0.889733	2.430	0.010989	80.964	< 2e-16
Workplace:55-59	1.961580	7.110	0.012588	155.833	< 2e-16
CoffeeShops:60-64	0.665376	1.950	0.014604	45.563	< 2e-16
Home:60-64	-0.867726	0.420	0.014684	-59.093	< 2e-16
PublicPlaces:60-64	0.512075	1.670	0.011931	42.918	< 2e-16
Workplace:60-64	1.432445	4.190	0.013775	103.991	< 2e-16
CoffeeShops:65+	0.487009	1.630	0.015322	31.785	< 2e-16
Home:65+	-0.106660	0.899	0.013502	-7.899	2.80e-15
PublicPlaces:65+	0.616809	1.850	0.012182	50.634	< 2e-16
Workplace:65+	0.891140	2.440	0.014814	60.156	< 2e-16

Table 3.17: **Fitting the Multinomial Model to the Female Population Stratified by Age Group**

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
(Intercept)	10.422400	-	0.005455	1910.571	< 2e-16
Coffee Shops	-0.079336	0.924	0.007872	-10.078	< 2e-16
Home	1.059695	2.890	0.006330	167.403	< 2e-16
Public Places	1.557499	4.750	0.006002	259.485	< 2e-16
Workplace	-0.010711	0.989	0.007735	-1.385	0.16617
15-19	0.176257	0.868	0.007396	23.830	< 2e-16
20-24	-0.089243	0.915	0.007893	-11.307	< 2e-16
25-29	0.176257	1.190	0.007396	23.830	< 2e-16
30-34	0.018428	1.020	0.007679	2.400	0.01641
35-39	0.120069	1.130	0.007493	16.023	< 2e-16
40-44	-0.190041	0.827	0.008108	-23.438	< 2e-16
45-49	-0.256972	0.773	0.008261	-31.108	< 2e-16
50-54	-1.378115	0.252	0.012158	-113.348	< 2e-16
55-59	-1.374344	0.253	0.012140	-113.209	< 2e-16
60-64	-1.590250	0.204	0.013256	-119.964	< 2e-16
65+	-1.624399	0.197	0.013446	-120.810	< 2e-16
CoffeeShops:15-19	0.112961	1.120	0.011377	9.929	< 2e-16
Home:15-19	0.022421	1.020	0.009273	2.418	0.01561
PublicPlaces:15-19	0.132159	1.140	0.008755	15.0952	< 2e-16
Workplace:15-19	0.032022	1.030	0.011301	2.834	0.00460
CoffeeShops:20-24	0.045687	1.050	0.011321	4.036	5.44e-05
Home:20-24	0.022560	1.020	0.009145	2.467	0.01363
PublicPlaces:20-24	0.045600	1.050	0.008667	5.262	1.43e-07
Workplace:20-24	0.206632	1.230	0.010914	18.932	< 2e-16
CoffeeShops:25-29	-0.354932	0.701	0.011200	-31.690	< 2e-16
Home:25-29	-0.136596	0.872	0.008656	-15.780	< 2e-16
PublicPlaces:25-29	-0.095931	0.909	0.008171	-11.741	< 2e-16
Workplace:25-29	-0.002297	0.998	0.010491	-0.219	0.82666
CoffeeShops:30-34	-0.019589	0.981	0.011110	-1.763	0.07789
Home:30-34	0.021481	1.020	0.008899	2.414	0.01579
PublicPlaces:30-34	0.077243	1.080	0.008423	9.171	< 2e-16
Workplace:30-34	-0.171932	0.842	0.011141	-15.432	< 2e-16

Table 3.18: Continuation of Table 23

Coefficient	Estimate	Odds Ratio	Std. Error	z value	Pr(> z)
CoffeeShops:35-39	-0.057053	0.945	0.010891	-5.239	1.62e-07
Home:35-39	-0.151241	0.860	0.008781	-17.224	< 2e-16
PublicPlaces:35-39	-0.008555	0.991	0.008248	-1.037	0.29965
Workplace:35-39	-0.112159	0.894	0.010774	-10.411	< 2e-16
CoffeeShops:40-44	0.248296	1.280	0.011329	21.916	< 2e-16
Home:40-44	0.069444	1.070	0.009364	7.416	1.21e-13
PublicPlaces:40-44	0.185536	1.200	0.008849	20.966	< 2e-16
Workplace:40-44	-0.007740	0.992	0.011510	-0.672	0.50130
CoffeeShops:45-49	-0.283697	0.753	0.012481	-22.731	< 2e-16
Home:45-49	-0.090067	0.914	0.009651	-9.332	< 2e-16
PublicPlaces:45-49	-0.174211	0.840	0.009174	-18.991	< 2e-16
Workplace:45-49	-0.235652	0.790	0.012147	-19.401	< 2e-16
CoffeeShops:50-54	0.082401	1.090	0.017255	4.775	1.79e-06
Home:50-54	0.374490	1.450	0.013648	27.439	< 2e-16
Public Places:50-54	0.456963	1.580	0.013032	35.064	< 2e-16
Workplace:50-54	0.783973	2.190	0.015244	51.427	< 2e-16
CoffeeShops:55-59	-0.077759	0.925	0.017811	-4.366	1.27e-05
Home:55-59	-0.041380	0.959	0.014148	-2.925	0.00345
PublicPlaces:55-59	0.152390	1.160	0.013226	11.522	< 2e-16
Workplace:55-59	0.222451	1.250	0.016512	13.472	< 2e-16
CoffeeShops:60-64	-0.129101	0.879	0.019691	-6.556	5.51e-11
Home:60-64	-0.144835	0.865	0.015637	-9.263	< 2e-16
PublicPlaces:60-64	0.166997	1.180	0.014423	11.579	< 2e-16
Workplace:60-64	0.248719	1.280	0.017912	13.885	< 2e-16
CoffeeShops:65+	0.678735	1.970	0.017203	39.454	< 2e-16
Home:65+	0.198684	1.220	0.015298	12.988	< 2e-16
PublicPlaces:65+	0.476483	1.610	0.014381	33.132	< 2e-16
Workplace:65+	0.663797	1.940	0.017014	39.015	< 2e-16

3.5 Discussion of Results

We fitted the Poisson regression model specified in Equation (3.1) to various combinations of categories involving smoking and each of the background characteristics: Age, place of exposure, and region. For the Poisson regression, two models were considered: the additive model (without the interaction term in Equation (3.2)), also referred to as the null model, and the multiplicative model (Equation (3.1)), known as the full model. The additive model, analogous to the chi-square test for independence, is useful for determining whether an association exists between the categories. On the other hand, the multiplicative model provides deeper insights into the nature of these relationships. Both models were fitted; however, for simplicity, only the complete results for the multiplicative model are reported. Furthermore, the multiplicative models presented in Tables 13-24 explored the suggested association between age and region by comparing smoking prevalence within the five exposure places. The analysis results, as presented in Tables 13-24, are discussed in the following sections.

Regions (Entire population)

In terms of smoking likelihood, smoking in coffee shops is about 3.60 times more likely than smoking in a car. There is no significant difference between the odds of smoking at home and in the car. Smoking is most prevalent in public places by about 5.43 times compared to smoking in cars. Workplaces are about 1.46 times more likely to

smoke than those in a car. The likelihood of smoking in Al-Baha is about 0.42 times less likely than in Aseer. Smoking in the Eastern region is about 0.65 times less likely than in Aseer. Smoking in Hail is about 0.14 times less likely than in Aseer. Smoking in Jazan is about 0.78 times less likely than in Aseer. Smoking in Al-Jouf is about 0.33 times less likely than in Aseer, while smoking in Al-Madinah is about 0.67 times less likely than in Aseer. Smoking in Makkah is about 3.73 times more likely than in Aseer. Smoking in Najran is about 0.24 times less likely than in Aseer. Smoking in the Northern Borders is about 0.27 times less likely than in Aseer. Smoking in Al-Qassem is about 0.31 times less likely than in Aseer. Smoking in Riyadh is about 3.03 times more likely than in Aseer. Smoking in Tabouk is about 1.94 times more likely than in Aseer. The odds of smoking in coffee shops among individuals in Al-Baha are approximately 0.23 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking at home among individuals in Al-Baha are approximately 1.64 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Baha are approximately 0.60 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Baha are approximately 0.58 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in the Eastern region are approximately 1.03 times higher than the odds of smoking in a car by someone in the

same region compared to individuals in Aseer. The odds of smoking at home among individuals in the Eastern region are approximately 6.89 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in the Eastern region are approximately 2.51 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in the Eastern region are approximately 3.94 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Hail are approximately 0.33 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Hail are approximately 4.92 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Hail are approximately 3.03 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Hail are approximately 2.61 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Jazan are approximately 0.25 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Jazan are approximately 1.77 times higher than the odds of smoking in a car by someone

in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Jazan are approximately 0.68 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Jazan are approximately 0.30 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Jouf are approximately 0.23 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Jouf are approximately 2.56 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Jouf are approximately 1.59 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Jouf are approximately 0.75 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Madinah are approximately 0.61 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Madinah are approximately 1.64 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Madinah are approximately 2.38 times higher than the odds of smoking in a car by someone in the same region com-

pared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Madinah are approximately 1.25 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Makkah are approximately 0.38 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Makkah are approximately 1.85 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Makkah are approximately 0.98 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Makkah are approximately 0.86 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Najran are approximately 1.19 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Najran are approximately 1.60 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Najran are approximately 1.21 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Najran are approximately 2.22 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Northern Borders are approximately 0.21 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Northern Borders are approximately 2.65 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Northern Borders are approximately 1.19 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Northern Borders are approximately 0.64 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Qaseem are approximately 0.44 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Qaseem are approximately 1.65 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Qaseem are approximately 2.38 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Qaseem are approximately 1.71 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Riyadh are approximately 0.28 times lower than the odds of smoking in a car by someone in the same region

compared to individuals in Aseer. The odds of smoking at home among individuals in Riyadh are approximately 1.71 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Riyadh are approximately 0.62 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Riyadh are approximately 1.06 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Tabouk are approximately 0.21 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Tabouk are approximately 1.48 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Tabouk are approximately 0.64 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Tabouk are approximately 0.41 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

Regions (Male population)

In terms of smoking likelihood, smoking in coffee shops is about 3.61 times more likely than smoking in a car. Smoking at home is about 0.87 times less likely than smoking in a car. Smoking is most prevalent in public places by about 5.36 times compared to

smoking in a car. Workplaces are about 1.69 times more likely to smoke than those in a car.

The likelihood of smoking in Al-Baha is about 0.41 times less likely than in Aseer. Smoking in the Eastern region is about 0.74 times less likely than in Aseer. Smoking in Hail is about 0.14 times less likely than in Aseer. There are no significant differences between the odds of smoking in a car among Jazan and Aseer. Smoking in Al-Jouf is about 0.39 times less likely than in Aseer while smoking in Al-Madinah is about 0.90 times less likely than in Aseer. Smoking in Makkah is about 4.16 times more likely than in Aseer. Smoking in Najran is about 0.28 times less likely than in Aseer. Smoking in Northern Borders is about 0.23 times less likely than in Aseer. Smoking in Al-Qassem is about 0.36 times less likely than in Aseer. Smoking in Riyadh is about 3.50 times more likely than in Aseer. Smoking in Tabouk is about 1.58 times more likely than in Aseer.

The odds of smoking in coffee shops among individuals in Al-Baha are approximately 0.31 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Baha are approximately 1.53 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Baha are approximately 0.77 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Baha are approximately 0.58 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in the eastern region are approximately 1.40 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in the Eastern region are approximately 5.85 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in the eastern region are approximately 2.90 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in the eastern region are approximately 4.42 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Hail are approximately 0.37 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Hail are approximately 4.36 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Hail are approximately 3.50 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Hail are approximately 2.35 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Jazan are approximately 0.30 times lower than the odds of smoking in a car by someone in the same region compared

to individuals in Aseer. The odds of smoking at home among individuals in Jazan are approximately 1.14 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Jazan are approximately 0.73 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Jazan are approximately 0.33 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Jouf are approximately 0.35 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Jouf are approximately 1.80 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Jouf are approximately 1.69 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Jouf are approximately 0.82 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Madinah are approximately 0.72 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Madinah are approximately 1.08 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds

of smoking in public places among individuals in Al-Madinah are approximately 2.15 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Madinah are approximately 1.24 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Makkah are approximately 0.53 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Makkah are approximately 1.48 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Makkah are approximately 1.11 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Makkah are approximately 0.94 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Najran are approximately 1.54 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Najran are approximately 1.38 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Najran are approximately 1.56 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Najran are approx-

imately 2.24 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Northern Borders are approximately 0.34 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Northern Borders are approximately 2.83 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Northern Borders are approximately 1.63 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Northern Borders are approximately 0.83 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Qaseem are approximately 0.54 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Qaseem are approximately 1.34 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Qaseem are approximately 3.03 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Qaseem are approximately 1.58 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Riyadh are approximately 0.36 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Riyadh are approximately 1.40 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Riyadh are approximately 0.74 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Riyadh are approximately 1.09 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Tabouk are approximately 0.51 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Tabouk are approximately 1.31 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Tabouk are approximately 0.93 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Tabouk are approximately 0.87 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

Regions (Female population)

In terms of smoking likelihood, smoking in coffee shops is about 3.59 times more likely than smoking in a car. Smoking at home is about 1.13 times less likely than smoking in a car. Smoking is most prevalent in public places by about 5.50 times compared to smoking in a car. Workplaces are about 1.23 times more likely to smoke than those in a car.

The likelihood of smoking in Al-Baha is about 0.43 times less likely than in Aseer. Smoking in the Eastern region is about 0.56 times less likely than in Aseer. Smoking in Hail is about 0.14 times less likely than in Aseer. Smoking in Jazan is about 0.56 times less likely than in Aseer. Smoking in Al-Jouf is about 0.28 times less likely than in Aseer, while smoking in Al-Madinah is about 0.46 times less likely than in Aseer. Smoking in Makkah is about 3.32 times more likely than in Aseer. Smoking in Najran is about 0.21 times less likely than in Aseer. Smoking in the Northern Borders is about 0.32 times less likely than in Aseer. Smoking in Al-Qassem is about 0.25 times less likely than in Aseer. Smoking in Riyadh is about 2.59 times more likely than in Aseer. Smoking in Tabouk is about 2.29 times more likely than in Aseer.

The odds of smoking in coffee shops among individuals in Al-Baha are approximately 0.16 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Baha are approximately 1.71 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Baha are approximately 0.45 times lower than

the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Baha are approximately 0.59 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in the Eastern region are approximately 0.56 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in the Eastern region are approximately 8.17 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in the Eastern region are approximately 2.04 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in the Eastern region are approximately 2.86 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Hail are approximately 2.28 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Hail are approximately 5.36 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Hail are approximately 2.57 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Hail are approximately 2.95 times

higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Jazan are approximately 0.17 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Jazan are approximately 2.79 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Jazan are approximately 0.60 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Jazan are approximately 0.19 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Jouf are approximately 0.06 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Jouf are approximately 3.47 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Jouf are approximately 1.45 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Jouf are approximately 0.58 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Madinah are approxi-

mately 0.40 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Madinah are approximately 2.64 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Madinah are approximately 2.85 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Madinah are approximately 1.04 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Makkah are approximately 0.19 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Makkah are approximately 2.25 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Makkah are approximately 0.84 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Makkah are approximately 0.69 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Najran are approximately 0.72 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Najran are approximately 1.88 times higher than the odds of smoking in a car by

someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Najran are approximately 0.76 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Najran are approximately 2.03 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Northern Borders are approximately 0.13 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Northern Borders are approximately 2.46 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Northern Borders are approximately 0.89 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Northern Borders are approximately 0.50 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Al-Qaseem are approximately 0.30 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Al-Qaseem are approximately 2.04 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Al-Qaseem are approximately 1.52

times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Al-Qaseem are approximately 1.83 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Riyadh are approximately 0.17 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at home among individuals in Riyadh are approximately 2.1 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Riyadh are approximately 0.47 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Riyadh are approximately 0.94 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer.

The odds of smoking in coffee shops among individuals in Tabouk are approximately 0.01 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking at Home among individuals in Tabouk are approximately 1.51 times higher than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in public places among individuals in Tabouk are approximately 0.45 times lower than the odds of smoking in a car by someone in the same region compared to individuals in Aseer. The odds of smoking in workplaces among individuals in Tabouk are approximately 0.02 times lower than the odds of smoking in a car by someone in the

same region compared to individuals in Aseer.

Comparison of Smoking Likelihood by Gender Across Regions

Males versus Females

Both males and females are similarly likely to smoke in coffee shops compared to cars; for males, the odds are about 3.61 times higher, while for females, the odds are slightly lower at 3.59 times higher. Smoking at home shows a notable difference between genders. For males, smoking at home is 0.87 times less likely than smoking in a car. For females, the odds are slightly higher, at 1.13 times less likely, and greater likelihood than males. Smoking in public places is the most prevalent for both genders, with females showing a slightly higher likelihood than males. The odds are 5.36 times higher for males and 5.50 times higher for females compared to smoking in a car. The odds of smoking in workplaces are higher for males (1.69 times) compared to females (1.23 times). This indicates that males are more likely to smoke in workplaces than females, relative to smoking in a car.

Smoking is less likely in Al-Baha compared to Aseer for both genders, with a similar reduction in odds. It is 0.41 times less likely for males, while for females, it is slightly lower at 0.43 times less likely. Smoking is less likely in the Eastern region for both genders, but females significantly reduce their odds. Males are 0.74 times less likely to smoke, while females are 0.56 times less likely. Both males and females have the

same likelihood of smoking in Hail compared to Aseer, at 0.14 times less likely. For males, there is no significant difference in smoking likelihood between Jazan and Aseer. However, females are 0.56 times less likely to smoke in Jazan than in Aseer. Smoking in Al-Jouf is less likely for both genders, but females show a more significant reduction in odds. Males are 0.39 times less likely, while females are 0.28 times less likely. Smoking is less likely in Al-Madinah for both genders. For males, it is 0.90 times less likely, while for females, it is 0.46 times less likely, indicating a more substantial reduction. Smoking is more likely in Makkah for both genders. However, males have higher odds (4.16 times more likely) than females (3.32 times more likely).

Smoking is less likely in Najran for both genders, with females showing slightly lower odds. Males are 0.28 times less likely, while females are 0.21 times less likely. Smoking in the Northern borders is less likely for both genders, with males being 0.23 times less likely and females being 0.32 times less likely. Smoking in Al-Qassem is less likely for both genders, with males being 0.36 times less likely and females being 0.25 times less likely, showing a more substantial reduction for females. Smoking in Riyadh is more likely for both genders, with males having higher odds (3.50 times more likely) compared to females (2.59 times more likely). Smoking in Tabouk is more likely for both genders, but females have higher odds. Males are 1.58 times more likely, while females are 2.29 times more likely. In Al-Baha, smoking in coffee shops is less likely for both genders, but the reduction is more substantial for females (0.16 times) than for males (0.31 times). Smoking at home in Al-Baha is more likely for both genders, with females showing slightly higher odds (1.71 times) than males (1.53 times).

In Al-Baha, smoking in public places is less likely for both genders, but there is a

more significant reduction for females (0.45 times) compared to males (0.77 times). Smoking in workplaces is less likely for both genders, but the odds are similar (0.58 times for males and 0.59 times for females). In the Eastern Region, smoking in coffee shops is more likely for males (1.4 times) but less likely for females (0.56 times). Smoking at home in the Eastern Region is significantly more likely for both genders, with females showing higher odds (8.17 times) compared to males (5.85 times). In the Eastern Region, smoking in public places is more likely for both genders, with males by about 2.9 times and females by about 2.04 times. In the Eastern Region, smoking in workplaces has more potential for both genders, with males 4.42 times and females 2.86 times.

In Hail, regarding smoking in coffee shops, males are less likely to smoke (0.37 times), while females are more likely (2.28 times). Smoking at home in Hail is more likely for both genders, with females showing higher odds (5.36 times) compared to males (4.36 times). In Hail, smoking in public places is more likely for both genders, with males at 3.5 times and females at 2.57 times. In Hail, smoking in workplaces is more likely for both genders, with males at 2.35 times and females at 2.95 times. In Jazan, smoking in coffee shops is less likely for both genders, with a more substantial reduction for females (0.17 times) compared to males (0.3 times). Smoking at home in Jazan is more likely for both genders, with females at 2.79 times and males at 1.14 times. In Jazan, smoking in public places is less likely for both genders, with females at 0.60 times and males at 0.73 times. In Jazan, smoking in workplaces is less likely for both genders, with females showing a much more substantial reduction (0.19 times) compared to males (0.33 times).

In Al-Jouf, smoking in coffee shops is less likely for both genders, with females showing a more substantial reduction (0.06 times) than males (0.35 times). Smoking at home in Al-Jouf is more likely for both genders, with females at 3.47 times and males at 1.8 times. In Al-Jouf, smoking in public places is more likely for both genders, with females at 1.45 times and males at 1.69 times. In Al-Jouf, smoking in workplaces is less likely for both genders, with males at 0.82 times and females at 0.58 times.

In Makkah, smoking in coffee shops is less likely for both genders, with females showing a more substantial reduction (0.19 times) than males (0.53 times). Smoking at home in Makkah is more likely for both genders, with females at 2.25 times and males at 1.48 times. In Makkah, smoking in public places is more likely for males (1.11 times) but less likely for females (0.84 times). In Makkah, smoking in Workplaces is less likely for both genders, with males at 0.94 times and females at 0.69 times. In Al-Madinah, smoking in coffee shops is less likely for both genders, but females show a more significant reduction (0.40 times lower) compared to males (0.72 times lower). Smoking at home in Al-Madinah is more likely for both genders, with females (2.64 times higher) showing a higher likelihood compared to males (1.08 times higher). In Al-Madinah, smoking in public places is more likely for both genders, with females (2.85 times higher) again showing greater odds compared to males (2.15 times higher). In Al-Madinah, smoking in workplaces is more likely for both genders, with males (1.24 times higher) showing slightly greater odds than females (1.04 times higher).

In Najran, smoking in coffee shops is more likely for males (1.54 times higher) but less likely for females (0.72 times lower). Smoking at home in Najran is more likely for both genders, with females (1.88 times higher) showing slightly greater odds compared

to males (1.38 times higher). In Najran, smoking in public places is more likely for males (1.56 times higher) but less likely for females (0.76 times lower). In Najran, smoking in workplaces is more likely for both genders, with males (2.24 times higher) showing slightly higher odds than females (2.03 times higher). In Northern Borders, smoking in coffee shops is less likely for both genders, but females (0.13 times lower) show a much more significant reduction than males (0.34 times lower). Smoking at home in Northern Borders is more likely for both genders, with males (2.83 times higher) showing greater odds than females (2.46 times higher). In Northern Borders, smoking in public places is more likely for males (1.63 times higher) but less likely for females (0.89 times lower). In Northern Borders, smoking in workplaces is less likely for both genders, with females (0.50 times lower) showing a more significant reduction compared to males (0.83 times lower). In Al-Qaseem, smoking in coffee shops is less likely for both genders, with females (0.30 times lower) showing a more significant reduction than males (0.54 times lower).

Smoking at home in Al-Qaseem is more likely for both genders, with females (2.04 times higher) showing more substantial odds compared to males (1.34 times higher). In Al-Qaseem, smoking in public places is more likely for both genders, with males (3.03 times higher) showing more substantial odds than females (1.52 times higher). In Al-Qaseem, smoking in workplaces is more likely for both genders, with females (1.83 times higher) showing slightly greater odds compared to males (1.58 times higher). In Riyadh, smoking in coffee shops is less likely for both genders, with females (0.17 times lower) showing a more significant reduction than males (0.36 times lower). Smoking at home in Riyadh is more likely for both genders, with females (2.10 times higher)

showing slightly greater odds compared to males (1.40 times higher). In Riyadh, smoking in public places is less likely for both genders, with females (0.47 times lower) showing a more significant reduction compared to males (0.74 times lower). In Riyadh, smoking in workplaces is less likely for females (0.94 times lower) but more likely for males (1.09 times higher).

In Tabouk, smoking in coffee shops is much less likely for both genders, with females (0.01 times lower) showing a much more substantial reduction than males (0.51 times lower). Smoking at home in Tabouk is more likely for both genders, with males (1.31 times higher) showing slightly greater odds than females (1.51 times higher). In Tabouk, smoking in public places is less likely for both genders, with females (0.45 times lower) showing a more significant reduction compared to males (0.93 times lower). In Tabouk, smoking in workplaces is much less likely for both genders, with females (0.02 times lower) showing a more substantial reduction compared to males (0.87 times lower).

Males versus the Entire population

Both males and the entire population show a significant preference for smoking in coffee shops, with very similar odds (3.60 vs. 3.61 times higher) compared to cars. For the whole population, there is no significant difference between smoking at home and in a car, whereas males are less likely to smoke at home (0.87 times lower) compared to cars. Both males and the entire population show a strong preference for smoking in public places, with very similar odds (5.43 vs. 5.36 times higher). Males are more likely to smoke in workplaces (1.69 times higher) than the entire population (1.46

times higher) compared to cars. The odds of the entire population being 0.42 times less likely to smoke in Al-Baha and males' odds being 0.41 times less likely to smoke in Al-Baha compared to Aseer. The odds of the entire population being 0.65 times less likely to smoke in Al-Baha and males odds being 0.74 times less likely to smoke in the Eastern region than in Aseer.

Both the entire population and males have the same odds of smoking in Hail (0.14 times less likely) compared to smoking in Aseer. Males show no significant difference between smoking in Jazan and Aseer, while the entire population shows a lower likelihood of smoking in Jazan (0.78 times less likely). Males (0.39 times less likely) and the entire population (0.33 times less likely) are less likely to smoke in Al-Jouf than in Aseer. Males (0.90 times less likely) and the entire population (0.67 times less likely) are less likely to smoke in Al-Madinah than in Aseer. Both Males (4.16 times less likely) and the entire population (3.73 times less likely) are more likely to smoke in Makkah compared to smoking in Aseer. Males (0.28 times less likely) and the entire population (0.24 times less likely) are less likely to smoke in Najran than in Aseer. Males (0.23 times less likely) and the entire population (0.27 times less likely) are less likely to smoke in Northern Borders than in Aseer. Males (0.36 times less likely) and the entire population (0.31 times less likely) are less likely to smoke in Al-Qassem than in Aseer. Males are more likely to smoke in Riyadh (3.50 times more likely), and the entire population (3.03 times more likely) is more likely to smoke in Riyadh compared to Aseer. Males are more likely to smoke in Tabouk (1.58 times more likely, and the entire population (1.94 times more likely) is likelier to smoke in Tabouk) than in Aseer. In Al-Baha, the odds of smoking in coffee shops are less likely

for both males by about 0.31 times and the entire population by about 0.23 times compared to smoking in cars in the Aseer region. In Al-Baha, the odds of smoking at home are more likely for both males by about 1.53 times and the entire population by about 1.64 times compared to smoking in cars in the Aseer region. In Al-Baha, the odds of smoking in public places are less likely for both males by about 0.77 times and the entire population by about 0.60 times compared to smoking in cars in the Aseer region.

In Al-Baha, the odds of smoking in workplaces are less likely for both males, about 0.58 times, and the entire population, about 0.58 times compared to smoking in cars in the Aseer region. In the Eastern region, the odds of smoking in coffee shops are more likely for both males by about 1.40 times and the entire population by about 1.03 times compared to smoking in cars in the Aseer region. In the Eastern region, the odds of smoking at home are more likely for both males by about 5.85 times and the entire population by about 6.89 times compared to smoking in cars in the Aseer region. In the Eastern region, the odds of smoking in public places are more likely for both males by about 2.90 times and the entire population by about 2.51 times compared to smoking in cars in the Aseer region. In the Eastern region, the odds of smoking in workplaces are more likely for both males by about 4.42 times and the entire population by about 3.94 times compared to smoking in cars in the Aseer region. In Hail, the odds of smoking in coffee shops are less likely for both males, about 0.37 times, and the entire population, about 0.33 times compared to smoking in cars in the Aseer region.

In Hail, the odds of smoking at home are more likely for both males by about 4.36

times and the entire population by about 4.92 times compared to smoking in cars in the Aseer region. In Hail, the odds of smoking in public places are more likely for both males by about 3.50 times and the entire population by about 3.03 times compared to smoking in cars in the Aseer region. In Hail, the odds of smoking in workplaces are more likely for both males by about 2.35 times and the entire population by about 2.61 times compared to smoking in cars in the Aseer region. In Jazan, the odds of smoking in coffee shops are less likely for both males by about 0.25 times and the entire population by about 0.30 times compared to smoking in cars in the Aseer region. In Jazan, the odds of smoking at home are more likely for both males by about 1.14 times and the entire population by about 1.77 times compared to smoking in cars in the Aseer region. In Jazan, the odds of smoking in public places are less likely for both males by about 0.73 times and the entire population by about 0.68 times compared to smoking in cars in the Aseer region.

In Jazan, the odds of smoking in workplaces are less likely for both males by about 0.30 times and the entire population by about 0.33 times compared to smoking in cars in the Aseer region. In Al-Jouf, the odds of smoking in coffee shops are less likely for both males by about 0.35 times and the entire population by about 0.23 times compared to smoking in cars in the Aseer region. In Al-Jouf, the odds of smoking at home are more likely for both males by about 1.80 times and the entire population by about 2.56 times compared to smoking in cars in the Aseer region. In Al-Jouf, the odds of smoking in public places are more likely for both males by about 1.69 times and the entire population by about 1.59 times compared to smoking in cars in the Aseer region. In Al-Jouf, the odds of smoking in workplaces are less likely for both

males by about 0.82 times and the entire population by about 0.75 times compared to smoking in cars in the Aseer region.

In Al-Madinah, the odds of smoking in coffee shops are less likely for both males by about 0.72 times and the entire population by about 0.61 times compared to smoking in cars in the Aseer region. In Al-Madinah, the odds of smoking at home are more likely for both males by about 1.08 times and the entire population by about 1.64 times compared to smoking in cars in the Aseer region. In Al-Madinah, the odds of smoking in public places are more likely for both males, about 2.15 times, and the entire population, about 2.38 times, compared to smoking in cars in the Aseer region. In Al-Madinah, the odds of smoking in workplaces are more likely for both males by about 1.24 times and the entire population by about 1.25 times compared to smoking in cars in the Aseer region.

In Makkah, the odds of smoking in coffee shops are less likely for both males by about 0.53 times and the entire population by about 0.38 times compared to smoking in cars in the Aseer region. In Makkah, the odds of smoking at home are more likely for both males by about 1.48 times and the entire population by about 1.85 times compared to smoking in cars in the Aseer region. In Makkah, the odds of smoking in public places are more likely for males, about 1.11 times, while the odds of smoking in public places are less likely for the entire population, about 0.98 times compared to smoking in cars in the Aseer region. In Makkah, the odds of smoking in workplaces are less likely for both males by about 0.94 times and the entire population by about 0.86 times compared to smoking in cars in the Aseer region.

In Najran, the odds of smoking in coffee shops are more likely for both males by about

1.54 times and the entire population by about 1.19 times compared to smoking in cars in the Aseer region. In Najran, the odds of smoking at home are more likely for both males by about 1.38 times and the entire population by about 1.60 times compared to smoking in cars in the Aseer region. In Najran, the odds of smoking in public places are more likely for both males by about 1.56 times and the entire population by about 1.21 times compared to smoking in cars in the Aseer region. In Najran, the odds of smoking in workplaces are more likely for both males, about 2.24 times, and the entire population, about 2.22 times compared to smoking in cars in the Aseer region.

In Northern Borders, the odds of smoking in coffee shops are less likely for both males by about 0.34 times and the entire population by about 0.21 times compared to smoking in cars in the Aseer region. In Northern Borders, the odds of smoking at home are more likely for both males by about 2.83 times and the entire population by about 2.65 times compared to smoking in cars in the Aseer region. In Northern Borders, the odds of smoking in public places are more likely for both males by about 1.63 times and the entire population by about 1.19 times compared to smoking in cars in the Aseer region. In Northern Borders, the odds of smoking in workplaces are less likely for both males by about 0.83 times and the entire population by about 0.64 times compared to smoking in cars in the Aseer region.

In Al-Qaseem, the odds of smoking in coffee shops are less likely for both males by about 0.54 times and the entire population by about 0.44 times compared to smoking in cars in the Aseer region. In Al-Qaseem, the odds of smoking at home are more likely for both males by about 1.34 times and the entire population by about 1.65 times compared to smoking in cars in the Aseer region. In Al-Qaseem, the odds of

smoking in public places are more likely for both males by about 3.03 times and the entire population by about 2.38 times compared to smoking in cars in the Aseer region. In Al-Qaseem, the odds of smoking in workplaces are more likely for both males by about 1.58 times and the entire population by about 1.71 times compared to smoking in cars in the Aseer region. In Riyadh, the odds of smoking in coffee shops are less likely for both males, about 0.36 times, and the entire population, about 0.28 times compared to smoking in cars in the Aseer region. In Riyadh, the odds of smoking at home are more likely for both males by about 1.40 times and the entire population by about 1.71 times compared to smoking in cars in the Aseer region.

In Riyadh, the odds of smoking in public places are less likely for both males by about 0.74 times and the entire population by about 0.62 times compared to smoking in cars in the Aseer region. In Riyadh, the odds of smoking in workplaces are more likely for both males by about 1.09 times and the entire population by about 1.06 times compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking in coffee shops are less likely for both males by about 0.51 times and the entire population by about 0.21 times compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking at home are more likely for both males by about 1.31 times and the entire population by about 1.48 times compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking in public places are less likely for both males by about 0.93 times and the entire population by about 0.64 times compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking in workplaces are less likely for both males about 0.87 times and the entire population about 0.41 times compared to smoking in cars in the Aseer region.

Females versus Entire population

Both females and the entire population show a significant preference for smoking in coffee shops, with very similar odds (3.60 vs. 3.59 times more likely) and negligible differences between them compared to cars. For the entire population, there is no difference between smoking at home and in a car, whereas females are more likely to smoke at home about 1.13 times compared to cars. Both males and the entire population show a strong preference for smoking in public places, with very similar odds (5.43 vs. 5.36 times higher). Both females (1.69 times higher) and the entire population (1.46 times higher) are more likely to smoke in workplaces than cars.

The odds of the entire population are 0.42 times less likely to smoke in Al-Baha, and the odds of females are 0.43 times less likely to smoke in Al-Baha than in Aseer. The odds of the entire population are 0.65 times less likely to smoke in Al-Baha, and the odds of females are 0.56 times less likely to smoke in the Eastern region than in Aseer. Both the entire population and females have the same odds of smoking in Hail, about 0.14 times less likely compared to smoking in Aseer. The odds of females are about 0.56 times less likely to smoke in Jazan than in Aseer, while the entire population shows a lower likelihood of about 0.78 times less likely compared to smoking in Jazan. The odds of females are about 0.28 times less likely to smoke in Al-Jouf, while the odds of the entire population are about 0.33 times less likely to smoke in Al-Jouf than in Aseer. The odds of females are about 0.46 times less likely to smoke in Al-Madinah, while the odds of the entire population are about 0.67 times less likely to smoke in Al-Madinah compared to smoking in Aseer. The odds of females are about 3.32 times

more likely to smoke in Makkah, while the odds of the entire population are about 3.73 times more likely to smoke in Makkah than in Aseer.

The odds of females are about 0.21 times less likely to smoke in Najran, while the odds of the entire population are about 0.24 times less likely to smoke in Najran than in Aseer. The odds of Females are about 0.32 times less likely to smoke in Northern Borders, while the odds of the entire population are about 0.27 times less likely to smoke in Northern Borders compared to smoking in Aseer. The odds of Females are about 0.25 times less likely to smoke in Al-Qassem, while the odds of the entire population are about 0.31 times less likely to smoke in Al-Qassem compared to smoking in Aseer. The odds of Females are about 2.59 times more likely to smoke in Riyadh, while the odds of the entire population are about 3.03 times more likely to smoke in Riyadh compared to smoking in Aseer.

The odds of females are about 2.29 times more likely to smoke in Tabouk, while the odds of the entire population are about 1.94 times more likely to smoke in Tabouk than in Aseer. In Al-Baha, the odds of smoking in coffee shops are less likely for both females, about 0.16 times, and the entire population, about 0.23 times compared to smoking in cars in the Aseer region. In Al-Baha, the odds of smoking at home are more likely for both females by about 1.71 times and the entire population by about 1.64 times compared to smoking in cars in the Aseer region. In Al-Baha, the odds of smoking in public places are less likely for both females, about 0.45 times, and for the entire population, about 0.60 times compared to smoking in cars in the Aseer region. In Al-Baha, the odds of smoking in workplaces are less likely for both females, about 0.59 times, and for the entire population, about 0.58 times compared to smoking in

cars in the Aseer region.

In the Eastern region, the odds of smoking in coffee shops are less likely for females, about 0.56 times. In contrast, the odds of smoking in coffee shops are more likely for the entire population, about 1.03 times compared to smoking in cars in the Aseer region. In the Eastern region, the odds of smoking at home are more likely for both females by about 8.17 times and the entire population by about 6.89 times compared to smoking in cars in the Aseer region. In the Eastern region, the odds of smoking in public places are more likely for both females about 2.04 times and the entire population about 2.51 times compared to smoking in cars in the Aseer region. In the Eastern region, the odds of smoking in workplaces are more likely for both females by about 2.86 times and the entire population by about 3.94 times compared to smoking in cars in the Aseer region.

In Hail, the odds of smoking in coffee shops are more likely for females, about 2.28 times, while the odds of smoking in coffee shops are less likely for the entire population, about 0.33 times compared to smoking in cars in the Aseer region. In Hail, the odds of smoking at home are more likely for both females by about 5.36 times and the entire population by about 4.92 times compared to smoking in cars in the Aseer region. In Hail, the odds of smoking in public places are more likely for both females by about 2.57 times and the entire population by about 3.03 times compared to smoking in cars in the Aseer region. In Hail, the odds of smoking in workplaces are more likely for both females by about 2.95 times and the entire population by about 2.61 times compared to smoking in cars in the Aseer region.

In Jazan, the odds of smoking in coffee shops are less likely for females by about 0.17

times and the entire population by about 0.30 times compared to smoking in cars in the Aseer region. In Jazan, the odds of smoking at home are more likely for both females by about 2.79 times and the entire population by about 1.77 times compared to smoking in cars in the Aseer region. In Jazan, the odds of smoking in public places are less likely for both females about 0.60 times and the entire population about 0.68 times compared to smoking in cars in the Aseer region. In Jazan, the odds of smoking in workplaces are less likely for both females by about 0.19 times and the entire population by about 0.33 times compared to smoking in cars in the Aseer region.

In Al-Jouf, the odds of smoking in coffee shops are less likely for females by about 0.06 times and for the entire population by about 0.23 times compared to smoking in cars in the Aseer region. In Al-Jouf, the odds of smoking at home are more likely for both females, about 3.47 times, and the entire population, about 2.56 times compared to smoking in cars in the Aseer region. In Al-Jouf, the odds of smoking in public places are more likely for both females by about 1.45 times and the entire population by about 1.59 times compared to smoking in cars in the Aseer region. In Al-Jouf, the odds of smoking in workplaces are less likely for both females by about 0.58 times and the entire population by about 0.75 times compared to smoking in cars in the Aseer region.

In Al-Madinah, the odds of smoking in coffee shops are less likely for both females, about 0.40 times, and the entire population, about 0.61 times compared to smoking in cars in the Aseer region. In Al-Madinah, the odds of smoking at home are more likely for both females by about 2.64 times and the entire population by about 1.64 times compared to smoking in cars in the Aseer region. In Al-Madinah, the odds of

smoking in public places are more likely for both females about 2.85 times and the entire population about 2.38 times compared to smoking in cars in the Aseer region. In Al-Madinah, the odds of smoking in workplaces are more likely for both females by about 1.04 times and the entire population by about 1.25 times compared to smoking in cars in the Aseer region. In Makkah, the odds of smoking in coffee shops are less likely for females by about 0.19 times and the entire population by about 0.38 times compared to smoking in cars in the Aseer region.

In Makkah, the odds of smoking at home are more likely for both females by about 2.25 times and the entire population by about 1.85 times compared to smoking in cars in the Aseer region. In Makkah, the odds of smoking in public places are less likely for both females by about 0.84 times and the entire population by about 0.98 times compared to smoking in cars in the Aseer region. In Makkah, the odds of smoking in workplaces are less likely for both females by about 0.69 times and the entire population by about 0.86 times compared to smoking in cars in the Aseer region. In Najran, the odds of smoking in coffee shops are less likely for females by about 0.72 times, while the odds of smoking in coffee shops are more likely for the entire population by about 1.19 times compared to smoking in cars in the Aseer region. In Najran, the odds of smoking at home are more likely for both females by about 1.88 times and the entire population by about 1.60 times compared to smoking in cars in the Aseer region.

In Najran, the odds of smoking in public places are less likely for females by about 0.76 times, while the odds of smoking in public places are more likely for the entire population by about 1.21 times compared to smoking in cars in the Aseer region. In

Najran, the odds of smoking in workplaces are more likely for both females by about 2.03 times and the entire population by about 2.22 times compared to smoking in cars in the Aseer region. In Northern Borders, the odds of smoking in coffee shops are less likely for both females by about 0.13 times and the entire population by about 0.21 times compared to smoking in cars in the Aseer region. In Northern Borders, the odds of smoking at home are more likely for both females by about 2.46 times and the entire population by about 2.65 times compared to smoking in cars in the Aseer region. In Northern Borders, the odds of smoking in public places are more likely for females, about 0.89 times. In contrast, the odds of smoking in public places are less likely for the entire population, about 1.19 times compared to smoking in cars in the Aseer region. In Northern Borders, the odds of smoking in workplaces are less likely for both females by about 0.50 times and the entire population by about 0.64 times compared to smoking in cars in the Aseer region.

In Al-Qaseem, the odds of smoking in coffee shops are less likely for both females about 0.30 times and the entire population about 0.44 times compared to smoking in cars in the Aseer region. In Al-Qaseem, the odds of smoking at home are more likely for both females by about 2.04 times and the entire population by about 1.65 times compared to smoking in cars in the Aseer region. In Al-Qaseem, the odds of smoking in public places are more likely for both females about 1.52 times and the entire population about 2.38 times compared to smoking in cars in the Aseer region. In Al-Qaseem, the odds of smoking in workplaces are more likely for both females by about 1.83 times and the entire population by about 1.71 times compared to smoking in cars in the Aseer region. In Riyadh, the odds of smoking in coffee shops are less

likely for females, about 0.17 times, and the entire population, about 0.28 times, compared to smoking in cars in the Aseer region. In Riyadh, the odds of smoking at home are more likely for both females by about 2.10 times and the entire population by about 1.71 times compared to smoking in cars in the Aseer region. In Riyadh, the odds of smoking in public places are less likely for both females, about 0.47 times, and the entire population, about 0.62 times compared to smoking in cars in the Aseer region. In Riyadh, the odds of smoking in workplaces are less likely for females, about 0.94 times, while the odds of smoking in workplaces are more likely for the entire population, about 1.06 times, compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking in coffee shops are less likely for both females by about 0.01 times and the entire population by about 0.21 times compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking at home are more likely for both females by about 1.51 times and the entire population by about 1.48 times compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking in public places are less likely for both females by about 0.45 times and the entire population by about 0.64 times compared to smoking in cars in the Aseer region. In Tabouk, the odds of smoking in workplaces are less likely for both females, about 0. times, and the entire population by about 0.41 times compared to smoking in cars in the Aseer region.

Age (Entire population)

The likelihood of smoking under the age group of 65+ is significantly lower by 0.24 times compared to 10-14 age groups, followed by 60-64 age groups by about 0.26 times

compared to 10-14 age groups, 55-54 age groups by about 0.31 times compared to 10-14 age groups, 50-54 age groups by about 0.37 times compared to 10-14 age groups, 45-49 age groups by about 0.81 times, 40-44 age groups by about 0.81 times, 30-34 age groups by about 1.11 times compared to 10-14 age groups, 35-39 age groups by about 1.17 times compared to 10-14 age groups. The likelihood is significantly higher among age groups 20-24 by 1.33 times compared to 10-14 age groups, followed by 15-19 age groups by about 1.27 times compared to 10-14 age groups, 25-29 age groups by about 1.24 times compared to 10-14 age groups. The likelihood of smoking in different places of exposure is significantly higher by about 5.08 times in public places compared to smoking in the car, followed by home by about 2.90 times, and coffee shops by about 0.92 times compared to smoking in the car. The likelihood is significantly lower, about 0.91 times, in the workplace compared to cars. The odds of smoking in coffee shops among individuals aged 15–19 are about 1.32 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 15–19 are approximately 0.69 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in public places among individuals aged 15–19 are about 0.92 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in the workplace among individuals aged 15–19 are approximately 0.88 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in coffee shops among individuals aged 20–24 are about 1.81 times higher

than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 20–24 are about 0.68 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 20–24 are about 0.99 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in the workplace among individuals aged 20–24 are approximately 1.56 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 25–29 are approximately 1.69 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 25–29 are about 0.79 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 25–29 are approximately 1.22 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in the workplace among individuals aged 25–29 are about 2.92 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 35–39 are approximately 2.15 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 35–39 are about 0.86 times lower than the odds of smoking in a car by someone

in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 35–39 are approximately 1.51 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in the workplace among individuals aged 35–39 are about 4.16 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 40–44 are approximately 2.46 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 40–44 are about 1.04 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 40–44 are approximately 1.93 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in the workplace among individuals aged 40–44 are about 4.54 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 45–49 are approximately 1.83 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 45–49 are about 0.85 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 45–49 are approximately 1.54 times higher than the odds of smoking in a car by someone in the same age group compared to

individuals aged 10–14. The odds of smoking in the workplace among individuals aged 45–49 are about 3.95 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 50–54 are approximately 2.68 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 50–54 are about 1.21 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 50–54 are approximately 2.32 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 50–54 are about 5.83 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 55–59 are about 2.32 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 55–59 are approximately 0.91 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 55–59 are about 1.99 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 55–59 are approximately 4.41 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 60–64 are about 1.51 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 60–64 are approximately 0.60 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 60–64 are about 1.51 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 60–64 are approximately 2.85 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 65+ are about 1.77 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 65+ are approximately 1.03 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 65+ are about 1.78 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 65+ are approximately 2.18 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14.

Age (Males Population)

The likelihood of smoking under the age group of 65+ is significantly lower by 0.29 times compared to 10-14 age groups, followed by 60-64 age groups by about 0.31 times compared to 10-14 age groups, 55-59 age groups by about 0.37 times compared to 10-14 age groups, 50-54 age groups by about 0.49 times compared to 10-14 age groups, 45-49 age groups by about 0.85 times, 40-44 age groups by about 0.91 times, 30-34 age groups by about 1.20 times compared to 10-14 age groups, 35-39 age groups by about 1.21 times compared to 10-14 age groups. The likelihood is significantly higher among age groups 20-24 by 1.74 times compared to 10-14 age groups, followed by 15-19 age groups by about 1.66 times compared to 10-14 age groups, 25-29 age groups by about 1.29 times compared to 10-14 age groups. The likelihood of smoking in different places of exposure is significantly higher by about 5.42 times in public places compared to smoking in the car, followed by home by about 2.91 times, and coffee shops by about 0.91 times compared to smoking in the car. The likelihood is significantly lower, about 0.83 times in the workplace, than in cars.

The odds of smoking in coffee shops among individuals aged 15–19 are approximately 1.43 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 15–19 are about 0.52 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 15–19 are approximately 0.79 times lower than the odds of smoking in a car by someone in the same age group compared to

individuals aged 10–14.

The odds of smoking in workplaces among individuals aged 15–19 are about 0.82 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in coffee shops among individuals aged 20–24 are approximately 2.22 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 20–24 are about 0.49 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 20–24 are approximately 0.94 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 20–24 are about 1.83 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 25–29 are approximately 2.62 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 25–29 are about 0.71 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 25–29 are approximately 1.46 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 25–29 are about 5.09 times higher than the odds of smoking in a car by someone in

the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 30–34 are approximately 2.87 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 30–34 are about 0.88 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 30–34 are approximately 1.61 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 30–34 are about 6.22 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 35–39 are approximately 3.28 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 35–39 are about 0.86 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 35–39 are approximately 1.92 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 35–39 are about 7.78 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 40–44 are approximately 3.54 times higher than the odds of smoking in a car by someone in the same age group

compared to individuals aged 10–14. The odds of smoking at home among individuals aged 40–44 are about 1.01 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 40–44 are approximately 2.50 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 40–44 are about 8.39 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 45–49 are approximately 1.81 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 45–49 are about 0.80 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 45–49 are approximately 2.09 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 45–49 are about 7.38 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 50–54 are approximately 3.51 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 50–54 are about 1.08 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking

in public places among individuals aged 50–54 are approximately 2.58 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 50–54 are about 8.31 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 55–59 are about 3.29 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 55–59 are approximately 0.88 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 55–59 are about 2.43 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 55–59 are approximately 7.11 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 60–64 are about 1.95 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 60–64 are approximately 0.42 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 60–64 are about 1.67 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged

60–64 are approximately 4.19 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 65+ are about 1.63 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 65+ are approximately 0.90 times lower than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 65+ are about 1.85 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 65+ are approximately 2.44 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14.

Age (Females Population)

The likelihood of smoking under the age group of 65+ is significantly lower by 0.197 times compared to 10-14 age groups, followed by 60-64 age groups by about 0.20 times compared to 10-14 age groups, 55-54 and 50-54 age groups have the same odds by about 0.25 times compared to 10-14 age groups, 45-49 age groups by about 0.77 times, 40-44 age groups by about 0.83 times, 15-19 age groups by about 0.87 times compared to 10-14 age groups, 20-24 age groups by about 0.92 times compared to 10-14 age groups. The likelihood is significantly higher among age groups 25-29 by about 1.19 times compared to 10-14 age groups, followed by 35-39 age groups by about 1.13 times compared to 10-14 age groups, 30-34 age groups by about 1.02 times compared

to 10-14 age groups. The likelihood of smoking in different places of exposure is significantly higher by about 4.75 times in public places compared to smoking in the car, followed by home by about 2.89 times, and coffee shops by about 0.92 times compared to smoking in the car. The likelihood is significantly lower, about 0.99 times, in the workplace than in cars.

The odds of smoking in coffee shops among individuals aged 15–19 are approximately 1.12 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 15–19 are about 1.02 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 15–19 are approximately 1.14 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in the workplace among individuals aged 15–19 are about 1.03 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 20–24 are approximately 1.05 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 20–24 are about 1.02 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 20–24 are approximately 1.05 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged

20–24 are about 1.23 times higher than the odds of smoking in a car by someone in the same age group, compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 25–29 are approximately 0.70 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 25–29 are about 0.87 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 25–29 are approximately 0.91 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 25–29 are the same as the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 30–34 are approximately 0.98 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 30–34 are about 1.02 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 30–34 are approximately 1.08 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 30–34 are about 0.84 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 35–39 are approximately

0.95 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 35–39 are about 0.86 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 35–39 are the same as the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 35–39 are about 0.89 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 40–44 are approximately 1.28 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 40–44 are about 1.07 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 40–44 are approximately 1.20 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 40–44 are the same as the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 45–49 are approximately 0.75 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 45–49 are about 0.91 times lower than the odds of smoking in a car by someone

in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 45–49 are approximately 0.84 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 45–49 are about 0.79 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 50–54 are approximately 1.09 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 50–54 are about 1.45 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 50–54 are approximately 1.58 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 50–54 are about 2.19 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 55–59 are about 0.93 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 55–59 are approximately 0.96 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 55–59 are about 1.16 times higher than the odds of smoking in a car by someone in the same age group compared to

individuals aged 10–14. The odds of smoking in workplaces among individuals aged 55–59 are approximately 1.25 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 60–64 are about 0.88 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 60–64 are approximately 0.87 times lower than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 60–64 are about 1.18 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 60–64 are approximately 1.28 times higher than the odds of smoking in a car by someone in the same age group compared to individuals aged 10–14.

The odds of smoking in coffee shops among individuals aged 65+ are about 1.97 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking at home among individuals aged 65+ are approximately 1.22 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in public places among individuals aged 65+ are about 1.61 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14. The odds of smoking in workplaces among individuals aged 65+ are approximately 1.94 times higher than those of smoking in a car by someone in the same age group compared to individuals aged 10–14.

Comparison of Smoking Likelihood by Gender

Across Age Groups

Males versus Females

For those under age 65+, the odds are less likely to smoke for both genders, with Males about 0.29 times and Females about 0.20 times compared to the 10–14 age group. For smokers under age 60–64, the odds are less likely to smoke for both genders, with Males about 0.31 times and females about 0.20 times compared to the 10–14 age group. For those smoking under the age of 55–59, the odds are less likely to smoke for both genders, with Males about 0.37 times and females about 0.25 times compared to the 10–14 age group. For smokers under age 50–54, the odds are less likely to smoke for both genders, with Males about 0.49 times and females about 0.25 times compared to the 10–14 age group. For those smoking under age 45–49, the odds are less likely to smoke for both genders, with Males about 0.85 times and females about 0.77 times compared to the 10–14 age group. Smoking under age 40–44, the odds are less likely to smoke for both genders, with Males about 0.91 times and females about 0.83 times compared to the 10–14 age group. For smokers under age 35–39, the odds are more likely to smoke for both genders, with Males about 1.21 times and females about 1.13 times compared to the 10–14 age group. For smokers under age 30–34, the odds are more likely to smoke for both genders, with Males about 1.20 times and females about 1.02 times compared to the 10–14 age group. For smokers under age 25–29, the odds are more likely to smoke for both genders, with Males about 1.29 times and females

about 1.19 times compared to the 10–14 age group. For those smoking under age 20–24, the odds of smoking among males is about 1.74 times more likely compared to those aged between 10–14 years old, while females are about 0.92 times less likely compared to those aged between 10–14 years. For those smoking under 15–19, the odds of smoking among Males are about 1.66 times more likely compared to those aged between 10–14 years, while females are about 0.87 times less likely compared to those aged between 10–14 years.

The odds of smoking in coffee shops are less likely for both genders; males are about 0.91 times, and females are about 0.92 times compared to smoking in cars. Smoking at home is more likely for both genders. The odds of males are about 2.91 times, and the odds of females are about 2.89 times compared to smoking in a car. The odds of smoking in public places are more likely for both genders; males are about 5.42 times, and females are about 4.75 times compared to smoking in the car. The odds of smoking in workplaces are less likely for both genders; males are about 0.83 times, and females are about 0.99 times compared to smoking in the car.

Under age 15–19, the odds of smoking in coffee shops are more likely for both males about 1.43 times, and females about 1.12 times compared to smoking in a car under age 10–14 years old. Under age 15–19, the odds of smoking at home are less likely for males about 0.52 times, and the odds of smoking at home are more likely for females about 1.02 times compared to smoking in a car under age 10–14 years old. Under age 15–19, the odds of smoking in public places are less likely for males about 0.79 times, and the odds of smoking in public places are more likely for females about 1.14 times compared to smoking in a car under ages 10–14 years old. Under age 15–19,

the odds of smoking in workplaces are less likely for males about 0.82 times, and the odds of smoking in workplaces are more likely for females about 1.03 times compared to smoking in a car under age 10-14 years old.

Under age 20-24, the odds of smoking in coffee shops are more likely for both males about 2.22 times, and females about 1.05 times compared to smoking in a car under ages 10-14 years old. Under age 20-24, the odds of smoking at home are less likely for males about 0.49 times, and the odds of smoking at home are more likely for females about 1.02 times compared to smoking in a car under age 10-14 years old. Under age 20-24, the odds of smoking in public places are less likely for males about 0.94 times, and the odds of smoking in public places are more likely for females about 1.05 times compared to smoking in a car under age 10-14 years old. Under age 20-24, the odds of smoking in workplaces are more likely for both males about 1.83 times, and females about 1.23 times compared to smoking in a car under ages 10-14 years old.

Under age 25-29, the odds of smoking in coffee shops are more likely for males about 2.62 times, and the odds of smoking in coffee shops are less likely for females about 0.70 times compared to smoking in a car under age 10-14 years old. Under those aged 25-29, the odds of smoking at home are less likely for both males, about 0.71 times, and females, about 0.87 times, compared to smoking in a car under ages 10-14 years old. Under age 25-29, the odds of smoking in public places are more likely for males about 1.46 times, and the odds of smoking in public places are less likely for females about 0.91 times compared to smoking in a car under age 10-14 years old. Under the ages 25-29, the odds of smoking in workplaces are more likely by about 5.09 times compared to smoking in a car 10-14 years old, and for females, there are no significant

differences between smoking in workplaces and smoking in a car under aged 10-14 years old.

Under age 30-34, the odds of smoking in coffee shops are more likely for males about 2.87 times, and the odds of smoking in coffee shops are less likely for females about 0.98 times compared to smoking in a car under age 10-14 years old. Under age 30-34, the odds of smoking at home are less likely for males about 0.88 times, and the odds of smoking at home are more likely for females about 1.02 times compared to smoking in a car under age 10-14 years old. Under age 30-34, the odds of smoking in public places are more likely for both males about 1.61 times, and females about 1.08 times compared to smoking in a car under ages 10-14 years old. Under age 30-34, the odds of smoking in workplaces are more likely for males about 6.22 times, and the odds of smoking in workplaces are less likely for females about 0.84 times compared to smoking in a car under age 10-14 years old.

Under age 35-39, the odds of smoking in coffee shops are more likely for males about 3.28 times, and the odds of smoking in coffee shops are less likely for females about 0.95 times compared to smoking in a car under age 10-14 years old. Under age 35-39, the odds of smoking at home are equally less likely for both males and females, about 0.86 times compared to smoking in a car under age 10-14 years old. Under ages 35-39, the odds of smoking in public places are more likely for males about 1.92 times compared to smoking in a car under those aged 10-14 years old, and for females, there are no significant differences between smoking in public places and smoking in a car under aged 10-14 years old about. Under age 35-39, the odds of smoking in workplaces are more likely for males about 7.78 times, and the odds of smoking in workplaces are

less likely for females about 0.89 times compared to smoking in a car under age 10-14 years old.

Under age 40-44, the odds of smoking in coffee shops are more likely for both males about 3.54 times, and females about 1.28 times compared to smoking in a car under ages 10-14 years old. Under ages 40-44, the odds of smoking at home are less likely for both males, about 1.01 times, and females, about 1.07 times, compared to smoking in a car under ages 10-14 years old. Under age 40-44, the odds of smoking in public places are more likely for both males about 2.50 times, and females about 1.20 times compared to smoking in a car under ages 10-14 years old. Under ages 40-44, the odds of smoking in workplaces are more likely for males about 8.39 times compared to smoking in a car under ages 10-14 years old, and for females, there are no significant differences between smoking in workplaces and smoking in a car under aged 10-14 years old.

Under age 45-49, the odds of smoking in coffee shops are more likely for males about 1.81 times, and the odds of smoking in coffee shops are less likely for females about 0.75 times compared to smoking in a car under age 10-14 years old. Under ages 45-49, the odds of smoking at home are less likely for both males, about 0.80 times, and females, about 0.91 times, compared to smoking in a car under ages 10-14 years old. Under age 45-49, the odds of smoking in public places are more likely for males about 2.09 times, and the odds of smoking in public places are less likely for females about 0.84 times compared to smoking in a car under age 10-14 years old. Under age 45-49, the odds of smoking in workplaces are more likely for males by about 7.38 times, and the odds of smoking in workplaces are more likely for females by about 0.79 times

compared to smoking in a car under age 10-14 years old.

Under age 50-54, the odds of smoking in coffee shops are more likely for both males about 3.51 times, and females about 1.09 times compared to smoking in a car under ages 10-14 years old. Under those aged 50-54, the odds of smoking at home are more likely for both males, about 1.08 times, and females, about 1.45 times, compared to smoking in a car under ages 10-14 years old. Under age 50-54, the odds of smoking in public places are more likely for both males about 2.58 times, and females about 1.58 times compared to smoking in a car under ages 10-14 years old. Under age 50-54, the odds of smoking in workplaces are more likely for both males about 8.31 times, and females about 2.19 times compared to smoking in a car under ages 10-14 years old.

Under age 55-59, the odds of smoking in coffee shops are more likely for males about 3.29 times, and the odds of smoking in coffee shops are less likely for females about 0.93 times compared to smoking in a car under age 10-14 years old. Under those aged 55-59, the odds of smoking at home are less likely for both males, about 0.88 times, and females, about 0.96 times, compared to smoking in a car under those aged 10-14 years old. Under age 55-59, the odds of smoking in public places are more likely for both males about 2.43 times, and females about 1.16 times compared to smoking in a car under age 10-14 years old. Under age 55-59, the odds of smoking in workplaces are more likely for both males about 7.11 times, and females about 1.25 times compared to smoking in a car under age 10-14 years old.

Under age 60-64, the odds of smoking in coffee shops are more likely for males about 1.95 times, and the odds of smoking in coffee shops are less likely for females about (0.88 times compared to smoking in a car under ages 10-14 years old. Under those

aged 60-64, the odds of smoking at home are less likely for both males, about 0.42 times, and females, about 0.87 times, compared to smoking in a car under those aged 10-14 years old. Under age 60-64, the odds of smoking in public places are more likely for both males about 1.67 times, and females about 1.18 times compared to smoking in a car under ages 10-14 years old. Under age 60-64, the odds of smoking in workplaces are more likely for both males about 4.19 times, and females about 1.28 times compared to smoking in a car under ages 10-14 years old.

Under age 65+, the odds of smoking in coffee shops are more likely for both males about 1.97 times, and females about 1.63 times compared to smoking in a car under age 10-14 years old. Under age 65+, the odds of smoking at home are less likely for males about 0.90 times, and the odds of smoking at home are more likely for females about 1.22 times compared to smoking in a car under ages 10-14 years old. Under age 65+, the odds of smoking in public places are more likely for both males, about 1.85 times, and females, about 1.61 times compared to smoking in a car under ages 10-14 years old. For those under age 65+, the odds of smoking in workplaces are more likely for both males about 2.44 times, and females about 1.94 times compared to smoking in a car under ages 10-14 years old.

Males versus the Entire population

For those under age 65+, the odds are less likely to smoke for both males about 0.29 times, and the entire population about 0.24 times compared to the 10–14 age group. For smokers under age 60–64, the odds are less likely to smoke for both males about 0.31 times, and the entire population about 0.26 times compared to the 10–14 age

group. For smokers under age 55–59, the odds are less likely to smoke for both males about 0.37 times, and the entire population about 0.31 times compared to the 10–14 age group. For smokers under age 50–54, the odds are less likely to smoke for males about 0.49 times, and the entire population about 0.37 times compared to the 10–14 age group. For smokers under age 45–49, the odds are less likely to smoke for males about 0.85 times, and the entire population about 0.81 times compared to the 10–14 age group. For smokers under age 40–44, the odds are less likely to smoke for males about 0.91 times, and the entire population about 0.87 times compared to the 10–14 age group. For smokers under age 35–39, the odds are more likely to smoke for both males about 1.21 times, and the entire population about 1.17 times compared to the 10–14 age group. For smokers under age 30–34, the odds are more likely to smoke for both males about 1.20 times, and the entire population about 1.11 times compared to the 10–14 age group.

For smokers under age 25–29, the odds are more likely to smoke for both males about 1.29 times, and the entire population about 1.24 times compared to the 10–14 age group. For those smoking under the age of 20–24, the odds of smoking are more likely for both males about 1.74 times, and the entire population is about 1.33 times compared to those aged between 10–14 years old. For those smoking under the age of 15–19, the odds of smoking are more likely for both genders, with males about 1.66 times and the entire population about 1.27 times compared to those aged between 10–14 years old.

The odds of smoking in coffee shops are less likely for both males, about 0.91 times, and the entire population is about 0.92 times compared to smoking in cars. The odds

of smoking at home are more likely for males, about 2.91 times, and the odds for the entire population are about 2.90 times the odds compared to smoking in a car. The odds of smoking in public places are more likely for both males at about 5.42 times, and the entire population is about 5.08 times compared to smoking in the car. The odds of smoking in workplaces are less likely for both males, about 0.83 times, and the entire population is about 0.91 times compared to smoking in the car

Under ages 15–19, the odds of smoking in coffee shops are more likely for both males by about 1.43 times and the entire population by about 1.32 times compared to smoking in a car under ages 10-14 years old. Under ages 15–19, the odds of smoking at home are less likely for both males by about 0.52 times and the entire population by about 0.69 times compared to smoking in a car under ages 10-14 years old. Under age 15–19, the odds of smoking in public places are less likely for both males by about 0.79 times and the entire population by about 0.92 times compared to smoking in a car under age 10-14 years old. Under those aged 15–19, the odds of smoking in workplaces are less likely for both males about 0.82 times, and the entire population about 0.88 times compared to smoking in a car under 10-14 years old.

Under ages 20-24, the odds of smoking in coffee shops are more likely for both males about 2.22 times, and the entire population about 1.81 times compared to smoking in a car under ages 10-14 years old. Under age 20-24, the odds of smoking at home are less likely for both males by about 0.49 times and the entire population by about 0.68 times compared to smoking in a car under ages 10-14 years old. Under ages 20-24, the odds of smoking in public places are less likely for both males by about 0.94 times and the entire population by about 0.99 times compared to smoking in a car under ages

10-14 years old. For those under the age of 20-24, the odds of smoking in workplaces are more likely for both males by about 1.83 times and the entire population by about 1.56 times compared to smoking in a car for those under the age of 10-14 years old.

Under those aged 25-29, the odds of smoking in coffee shops are more likely for both males by about 2.62 times and the entire population by about 1.69 times compared to smoking in a car under age 10-14 years old. Under age 25-29, the odds of smoking at home are less likely for both males by about 0.71 times and the entire population by about 0.79 times compared to smoking in a car under ages 10-14 years old. Under age 25-29, the odds of smoking in public places are more likely for both males by about 1.46 times and the entire population by about 1.22 times compared to smoking in a car under ages 10-14 years old. Under those aged 25-29, the odds of smoking in workplaces are more likely for both males about 5.09 times, and the entire population about 2.94 times compared to smoking in a car under age 10-14 years old.

Under those aged 30-34, the odds of smoking in coffee shops are more likely for both males about 2.87 times, and the entire population about 2.00 times compared to smoking in a car under 10-14 years old. Under ages 30-34, the odds of smoking at home are less likely for both males by about 0.88 times and the entire population by about 0.94 times compared to smoking in a car under ages 10-14 years old. Under ages 30-34, the odds of smoking in public places are more likely for both males by about 1.61 times and the entire population by about 1.40 times compared to smoking in a car under ages 10-14 years old. Under age 30-34, the odds of smoking in workplaces are more likely for both males about 6.22 times, and the entire population about 3.50 times compared to smoking in a car under ages 10-14 years old.

Under ages 35-39, the odds of smoking in coffee shops are more likely for both males about 3.28 times, and the entire population about 2.15 times compared to smoking in a car under ages 10-14 years old. Under age 35-39, the odds of smoking at home are equally less likely for both males and the entire population, about 0.86 times compared to smoking in a car under ages 10-14 years old. Under those aged 35-39, the odds of smoking in public places are more likely for both males about 1.92 times, and the entire population about 1.51 times compared to smoking in a car under age 10-14 years old. Under age 35-39, the odds of smoking in workplaces are more likely for both males about 7.78 times, and the entire population about 4.16 times compared to smoking in a car under ages 10-14 years old.

Under those aged 40-44, the odds of smoking in coffee shops are more likely for both males about 3.54 times, and the entire population about 2.46 times compared to smoking in a car under age 10-14 years old. Under age 40-44, the odds of smoking at home are more likely for both males about 1.01 times, and the entire population about 1.04 times compared to smoking in a car under ages 10-14 years old. Under ages 40-44, the odds of smoking in public places are more likely for both males by about 2.50 times and the entire population by about 1.93 times compared to smoking in a car under ages 10-14 years old. Under age 40-44, the odds of smoking in workplaces are more likely for both males about 8.39 times, and the entire population about 4.54 times compared to smoking in a car under ages 10-14 years old.

Under ages 45-49, the odds of smoking in coffee shops are more likely for both males about 1.81 times, and the entire population about 1.83 times compared to smoking in a car under ages 10-14 years old. Under those aged 45-49, the odds of smoking at

home are less likely for both males about 0.80 times, and the entire population about 0.85 times compared to smoking in a car under 10-14 years old. Under ages 45-49, the odds of smoking in public places are more likely for both males about 2.09 times, and the entire population about 1.54 times compared to smoking in a car under ages 10-14 years old. Under the age 45-49, the odds of smoking in workplaces are more likely for both males about 7.38 times, and the entire population about 3.95 times compared to smoking in a car under the age of 10-14 years old.

Under ages 50-54, the odds of smoking in coffee shops are more likely for both males about 3.51 times, and the entire population about 2.68 times compared to smoking in a car under ages 10-14 years old. Under age 50-54, the odds of smoking at home are more likely for both males about 1.08 times, and the entire population about 1.21 times compared to smoking in a car under ages 10-14 years old. Under ages 50-54, the odds of smoking in public places are more likely for both males by about 2.58 times and the entire population by about 2.32 times compared to smoking in a car under ages 10-14 years old. Under age 50-54, the odds of smoking in workplaces are more likely for both males about 8.31 times, and the entire population about 5.83 times compared to smoking in a car under ages 10-14 years old.

Under ages 55-59, the odds of smoking in coffee shops are more likely for both males about 3.29 times, and the entire population about 2.32 times compared to smoking in a car under ages 10-14 years old. Under age 55-59, the odds of smoking at home are less likely for both males about 0.88 times, and the entire population about 0.91 times compared to smoking in a car under ages 10-14 years old. Under ages 55-59, the odds of smoking in public places are more likely for both males about 2.43 times,

and the entire population about 1.99 times compared to smoking in a car under ages 10-14 years old. Under age 55-59, the odds of smoking in workplaces are more likely for both males about 7.11 times, and the entire population about 4.41 times compared to smoking in a car under ages 10-14 years old.

Under those aged 60-64, the odds of smoking in coffee shops are more likely for both males by about 1.95 times and the entire population by about 1.51 times compared to smoking in a car under age 10-14 years old. Under those aged 60-64, the odds of smoking at home are less likely for both males about 0.42 times, and the entire population about 0.60 times compared to smoking in a car under age 10-14 years old. Under those aged 60-64, the odds of smoking in public places are more likely for both males about 1.67 times, and the entire population about 1.51 times compared to smoking in a car under age 10-14 years old. Under age 60-64, the odds of smoking in workplaces are more likely for both males about 4.19 times, and the entire population about 2.85 times compared to smoking in a car under ages 10-14 years old.

Under those aged 65+, the odds of smoking in coffee shops are more likely for both males about 1.97 times, and the entire population about 1.77 times compared to smoking in a car under age 10-14 years old. Under age 65+, the odds of smoking at home are less likely for males by about 0.90 times, and the odds of smoking at home are more likely for the entire population by about 1.03 times compared to smoking in a car for those under age 10-14 years old. Under those aged 65+, the odds of smoking in public places are more likely for both males about 1.85 times, and the entire population about 1.78 times compared to smoking in a car under age 10-14 years old. Under age 65+, the odds of smoking in workplaces are more likely for both

males by about 2.44 times and the entire population by about 2.18 times compared to smoking in a car under ages 10-14 years old.

Females versus the Entire Population

For those smoking under age 65+, the odds are less likely to smoke for both females by about 0.20 times, and the entire population by about 0.24 times compared to the 10-14 age group. For smokers under age 60-64, the odds are less likely to smoke for both females about 0.20 times, and the entire population about 0.26 times compared to the 10-14 age group. For smokers under age 55-59, the odds are less likely to smoke for both females about 0.25 times, and the entire population about 0.31 times compared to the 10-14 age group. For smokers under age 50-54, the odds are less likely to smoke for both females about 0.25 times, and the entire population about 0.37 times compared to the 10-14 age group. For smokers under age 45-49, the odds are less likely to smoke for both females about 0.77 times, and the entire population about 0.81 times compared to the 10-14 age group. For those smoking under age 40-44, the odds are less likely to smoke for both females about 0.83 times, and the entire population about 0.87 times compared to the 10-14 age group. For smokers under age 35-39, the odds are more likely to smoke for both females about 1.13 times, and the entire population about 1.17 times compared to the 10-14 age group.

For smokers under age 30-34, the odds are more likely to smoke for both females about 1.02 times, and the entire population about 1.11 times compared to the 10-14 age group. For smokers under age 25-29, the odds are more likely to smoke for both females about 1.19 times, and the entire population about 1.24 times compared to the

10–14 age group. For those smoking under the age of 20–24, the odds of smoking are less likely for females by about 0.92 times compared to those aged 10–14 years old, and the odds of smoking are more likely for the entire population by about 1.33 times compared to those aged between 10–14 years old. For those smoking under age 15–19, the odds of smoking are less likely for females by about 0.87 times, and the odds of smoking are more likely for the entire population by about 1.27 times compared to those aged between 10–14 years old.

The odds of Smoking in coffee shops are equally less likely for both females and the entire population, about 0.92 times compared to smoking in cars. Smoking at home is more likely for both females, about 2.89 times, and the entire population, about 2.90 times, compared to smoking in a car. The odds of Smoking in public places are more likely for females by about 4.75 times and the entire population by about 5.08 times compared to smoking in the car. The odds of Smoking in workplaces are less likely for both females, about 0.99 times, and the entire population by about 0.91 times compared to smoking in the car.

Under ages 15–19, the odds of smoking in coffee shops are more likely for both females by about 1.12 times and the entire population by about 1.32 times compared to smoking in a car under ages 10–14 years old. Under age 15–19, the odds of smoking at home are more likely for females by about 1.02 times, and the odds of smoking at home are less likely for the entire population by about 0.69 times compared to smoking in a car for those under 10–14 years old. Under ages 15–19, the odds of smoking in public places are more likely for females by about 1.14 times, and the odds of smoking in public places are less likely for the entire population by about 0.92 times compared

to smoking in a car under age 10-14 years old. For those under the age of 15–19, the odds of smoking in workplaces are more likely for females by about 1.03 times, and the odds of smoking in workplaces are less likely for the entire population by about 0.88 times compared to smoking in a car for those under the age of 10-14 years old.

Under those aged 20-24, the odds of smoking in coffee shops are more likely for both females by about 1.05 times and the entire population by about 1.81 times compared to smoking in a car under those aged 10-14 years old. Under age 20-24, the odds of smoking at home are more likely for females by about 1.03 times, and the odds of smoking at home are less likely for the entire population by about 0.68 times compared to smoking in a car under age 10-14 years old. Under age 20-24, the odds of smoking in public places are more likely for females by about 1.05 times, and the odds of smoking in public places are less likely for the entire population by about 0.99 times compared to smoking in a car under those aged 10-14 years old. Under the age of 20-24, the odds of smoking in workplaces are more likely for both females by about 1.23 times and the entire population by about 1.56 times compared to smoking in a car among those aged 10-14 years old.

Under the age of 25-29, the odds of smoking in coffee shops are less likely for females by about 0.70 times, and the odds of smoking in coffee shops are more likely for the entire population by about 1.69 times compared to smoking in a car for under age 10-14 years old. Under those aged 25-29, the odds of smoking at home are less likely for both females by about 0.87 times and the entire population by about 0.79 times compared to smoking in a car under those aged 10-14 years old. Under age 25-29, the odds of smoking in public places are less likely for females, about 0.91 times, and the

odds of smoking in public places are more likely for the entire population, about 1.22 times, compared to smoking in a car under age 10-14 years old. Under ages 25-29, the odds for females, there is no significant difference between smoking in workplaces and smoking in a car under ages 10-14 years old, and the odds of smoking in workplaces are more likely for the entire population about 2.94 times compared to smoking in a car under aged 10-14 years old.

Under those aged 30-34, the odds of smoking in coffee shops are less likely for females by about 0.98 times, and the odds of smoking in coffee shops are more likely for the entire population by about 2.00 times compared to smoking in a car under age 10-14 years old. Under age 30-34, the odds of smoking at home are more likely for females by about 1.02 times, and the odds of smoking at home are less likely for the entire population by about 0.94 times compared to smoking in a car for those under 10-14 years old. Under those aged 30-34, the odds of smoking in public places are more likely for both females by about 1.08 times and the entire population by about 1.40 times compared to smoking in a car under those aged 10-14 years old. Under the age of 30-34, the odds of smoking in workplaces are less likely for females by about 0.84 times, and the odds of smoking in workplaces are more likely for the entire population by about 3.50 times compared to smoking in a car under age 10-14 years old.

Under those aged 35-39, the odds of smoking in coffee shops are less likely for females by about 0.95 times, and the odds of smoking in coffee shops are more likely for the entire population by about 2.15 times compared to smoking in a car under age of 10-14 years old. Under age 35-39, the odds of smoking at home are equally less likely for both females and the entire population, about 0.86 times compared to smoking

in a car under age 10-14 years old. Under age 35-39, the odds of smoking in public places are less likely for females by about 0.99 times, and the odds of smoking in public places are more likely for the entire population by about 1.51 times compared to smoking in a car under age 10-14 years old. Under ages 35-39, the odds of smoking in workplaces are less likely for females by about 0.89 times, and the odds of smoking in workplaces are more likely for the entire population by about 4.16 times compared to smoking in a car under age 10-14 years old.

Under those aged 40-44, the odds of smoking in coffee shops are more likely for both females by about 1.28 times and the entire population by about 2.46 times compared to smoking in a car under age 10-14 years old. Under those aged 40-44, the odds of smoking at home are more likely for both females about 1.07 times, and the entire population about 1.04 times, compared to smoking in a car under age 10-14 years old.

Under those aged 40-44, the odds of smoking in public places are more likely for both females by about 1.20 times and the entire population by about 1.93 times compared to smoking in a car under those aged 10-14 years old. Under age 40-44, the odds of smoking in workplaces are less likely for females by about 0.99 times, and the odds of smoking in workplaces are more likely for the entire population by about 4.54 times compared to smoking in a car under age 10-14 years old.

Under age 45-49, the odds of smoking in coffee shops are less likely for females by about 0.75 times, and the odds of smoking in coffee shops are more likely for the entire population by about 1.83 times compared to smoking in a car under age 10-14 years old. Under ages 45-49, the odds of smoking at home are less likely for both females by about 0.91 times and the entire population by about 0.85 times compared

to smoking in a car under ages 10-14 years old. Under age 45-49, the odds of smoking in public places are more likely for females by about 0.84 times, and the odds of smoking in public places are more likely for the entire population by about 1.54 times compared to smoking in a car under age 10-14 years old. Under those aged 45-49, the odds of smoking in workplaces are less likely for females by about 0.79 times, and the odds of smoking in workplaces are more likely for the entire population by about 3.95 times compared to smoking in a car under those aged 10-14 years old.

Under those aged 50-54, the odds of smoking in coffee shops are more likely for both females by about 1.09 times and the entire population by about 2.68 times compared to smoking in a car under those aged 10-14 years old. Under those aged 50-54, the odds of smoking at home are more likely for both females by about 1.45 times and the entire population by about 1.21 times compared to smoking in a car under those aged 10-14 years old. Under age 50-54, the odds of smoking in public places are more likely for both females by about 1.58 times and the entire population by about 2.32 times compared to smoking in a car under those aged 10-14 years old. Under those aged 50-54, the odds of smoking in workplaces are more likely for both females by about 2.19 times and the entire population by about 5.83 times compared to smoking in a car under those aged 10-14 years old.

Under age 55-59, the odds of smoking in coffee shops are less likely for females, about 0.93 times, and the odds of smoking in coffee shops are more likely for the entire population, about 2.32 times compared to smoking in a car under age 10-14 years old. Under ages 55-59, the odds of smoking at home are less likely for both females by about 0.96 times and the entire population by about 0.91 times compared to smoking

in a car under ages 10-14 years old. Under age 55-59, the odds of smoking in public places are more likely for both females by about 1.16 times and the entire population by about 1.99 times compared to smoking in a car under those aged 10-14 years old. Among the ages 55-59, the odds of smoking in workplaces are more likely for both females by about 1.25 times and the entire population by about 4.41 times compared to smoking in a car under age 10-14 years old.

Under ages 60-64, the odds of smoking in coffee shops are less likely for females, about 0.88 times, and the odds of smoking in coffee shops are more likely for the entire population, about 1.51 times compared to smoking in a car under age 10-14 years old. Under those aged 60-64, the odds of smoking at home are less likely for both females by about 0.87 times and the entire population by about 0.60 times compared to smoking in a car under those aged 10-14 years old. Under age 60-64, the odds of smoking in public places are more likely for both females by about 1.18 times and the entire population by about 1.51 times compared to smoking in a car under those aged 10-14 years old. Under age 60-64, the odds of smoking in workplaces are more likely for both females by about 1.28 times and the entire population by about 2.85 times compared to smoking in a car under the age of 10-14 years old.

For those under age 65+, the odds of smoking in coffee shops are more likely for both females by about 1.97 times and the entire population by about 1.77 times compared to smoking in a car for those under 10-14 years old. Under the age of 65+, the odds of smoking at home are more likely for both females by about 1.22 times and the entire population by about 1.03 times compared to smoking in a car under 10-14 years old.

For those aged 65+, the odds of smoking in public places are more likely for both

females, about 1.61 times, and the entire population, about 1.78 times, compared to smoking in a car under age 10-14 years old. For those aged 65+, the odds of smoking in workplaces are more likely for both females by about 1.94 times and the entire population by about 2.18 times compared to smoking in a car under those aged 10-14 years old.

CHAPTER 4

HURDLE MODEL

4.1 Introduction

In real-world applications, data confined to the restricted domain of $[0,1)$ are prevalent in various disciplines, representing proportions, standardized values, or transformed metrics. Examples include demographic ratios (e.g., sex ratio, dependency ratio), economic indicators (e.g., savings-to-income ratio, capital-to-output ratio), and engineering metrics (e.g., aspect ratio in aerospace, extinction ratio in telecommunications) [41]. However, the diversity and variability of data sets in fields such as engineering, insurance, and medical sciences pose significant challenges, as no single distribution can universally capture the characteristics of all types of data [42], [43]. This limitation underscores the need to continuously develop new distributions to address the unique variations observed in diverse data sets. Distributions are indispensable tools for data scientists [44], [45] and other data users, allowing them to perform inferential analyses and extract insights to inform decision-making processes. These tools enable domain experts to interpret, characterize, and predict patterns in real-world

phenomena that generate the data.

Continuous distributions with bounded domains, such as the Beta distribution (Pearson type IV), are essential for modeling bounded phenomena like proportions or other characteristics confined to specific ranges. The Beta distribution, with origins tracing back to a 1676 correspondence between Sir Isaac Newton and Henry Oldenberg, remains one of the most extensively studied and widely recognized distributions in statistical theory. Despite its prominence, the Beta distribution is not universally applicable, necessitating the development of alternative or complementary distributions to model bounded data effectively. This ongoing effort reflects the critical role of statistical distributions in addressing the complexities and variations inherent in real-world data in diverse fields [46]. Numerous anomalies develop data expressed as percentages, rates, ratios, or proportions, and such data generally belong to specific intervals, most commonly the unit interval $(0, 1)$. Such bounded data often arise in diverse fields, including hydrology [47], actuarial science [48], political science [49], finance [50], psychology [51], and municipal waste management [52].

Various probabilistic models have been developed to model data in the interval $(0, 1)$, with the beta distribution being the most utilized. Despite its extensive application, the beta distribution has a significant limitation: its cumulative distribution function (CDF) is expressed entirely in special functions, making it a systematically intensive and sophisticated associated statistical inference. Several variant distributions have been proposed to address this, among which the Kumaraswamy distribution [53] is particularly notable. Although the Kumaraswamy distribution is beta-type and shares similarities with the beta distribution, it offers greater docility, as highlighted by Jones

[54], who compared the two distributions comprehensively.

In addition to the Kumaraswamy distribution, many other unit distributions have been developed to model data sets confined to the unit interval $[0, 1]$. These data sets often arise from various and complex anomalies, requiring distributions with greater flexibility and fitting expertise. Although the Kumaraswamy distribution is widely recognized, various variant distributions existed before its introduction, offering viable alternatives when the beta or Kumaraswamy distributions fail to achieve a satisfactory fit. Advanced advancements have introduced additional options to the beta distribution, including the unit upper truncated Weibull distribution [55], the Unit-Logistic distribution [56], the Topp-Leone distribution [57], Johnson $S'B$ distribution [58], Johnson S_B distribution [59], the two-parameter unit logistic distribution [60], the two-parameter unit-Birnbaum–Saunders distribution [61], the unit-Gompertz distribution [62], the unit-inverse Gaussian distribution [63], the unit-Rayleigh distribution [64], the Arcsine distribution [65], the Simplex distribution [66], the Reflected Generalized Topp-Leone distribution [67], the Beta Power distribution [68], the Exponentiated Kumaraswamy distribution [69], and the Marshall-Olkin Extended Kumaraswamy distribution [70], the unit-Lindley distribution [71], [72], the unit Burr-XII distribution [73], the unit log-log distribution [74], and flexible univariate and bivariate families of distributions for the unit interval. It is essential to acknowledge that many distributions discussed above involve partial gamma functions, significantly complicating their computational implementation. Furthermore, most of these distributions are characterized by more than two parameters, and their ranges often depend on specific parameter values, making parameter estimation challenging and complicated.

Despite these complexities, various distributions with bounded support have been derived from standard distributions through various mathematical transformations. Similarly, unit distributions can be developed through transformations of random variables. For example, if a random variable X is defined in the interval $0, \infty$, the transformation $Y = \exp(-X)$ can produce a unit distribution. Examples of this include the unit Rayleigh distribution and the log WE distribution, the Unit Gamma distribution, the Kumaraswamy distribution with parameters α and β , the Log-Lindley distribution with parameters α and β , the unit G/G distribution, unit Fréchet distribution, the unit Birnbaum-Saunders, which arise when X follows the Rayleigh distribution or the weighted exponential (WE) distribution [75], and Gamma distribution [76], Exponential-Exponential distribution [53], Lindley distribution with parameters α and β [77], the gamma/Gompertz (G/G) distribution by [78], [79] the Fréchet distribution [80], the Birnbaum Saunders distribution [70], [81] respectively.

Another transformation involves the random variable $Y = X/(1 + X)$, which follows the unit Lindley distribution [62] if X is distributed according to the Lindley distribution [73]. Suppose that the random variable X follows the improved second-degree Lindley (ISDL) distribution by [82], [72] defined a unit random variable as $Y = X/(1 + X)$, and the distribution of Y is called the unit ISDL distribution. If X follows a random variable, then $Y = 1 - \frac{1}{1 + \exp(\frac{X - \alpha}{\beta})}$. Follows a Johnson SB distribution with parameters α and β [59], the Johnson SB distribution [58], and the Unit Logistic distribution [56]. The distribution of Y is referred to as a standard normal distribution, a standard Laplace distribution, and a standard logistic distribution, respectively. Additionally, for $\lambda > 0$, the power transformation $Y = X^\lambda$ results in the

two-parameter Topp-Leone distribution [83] if X follows the Topp-Leone distribution [57].

4.2 The Proposed Model

The two-parameter Gompertz (G) distribution, introduced by [84], represents an extension of the exponential distribution and plays a crucial role in survival analysis. It is beneficial for constructing accurate actuarial and human mortality tables. The Gompertz distribution is a valuable model for survival distributions described by an increasing hazard rate. It is also widely used by demographers and actuaries to outline the distribution of adult life spans, as highlighted in [85]. Numerous studies have supported the statistical characterization and methodological advancements of the Gompertz distribution, as evidenced by the works of [86], [87], [88], [89], [90], [91], [92]. Inverse distributions have been incorporated into modeling frameworks within demography, biological studies, and actuarial research, as discussed in [92], [93], [94], [95], [96]. Specifically, the inverse Gompertz (IG) distribution was introduced by [97] as a lifetime model. Let X be a random variable following an IG distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. The cumulative distribution function (CDF) of X is given by:

$$F_X(x; \alpha, \beta) = \exp \left(\frac{\alpha}{\beta} \left(1 - \exp \left(\frac{\beta}{x} \right) \right) \right), \quad x > 0; \quad \alpha, \beta > 0. \quad (4.1)$$

Similarly, the probability density function (PDF) of X is formulated as:

$$f_X(x; \alpha, \beta) = \frac{\alpha}{x^2} \exp \left(\frac{\alpha}{\beta} \left(1 - \exp \left(\frac{\beta}{x} \right) \right) + \frac{\beta}{x} \right), \quad x > 0; \quad \alpha, \beta > 0. \quad (4.2)$$

[98] introduced a novel existence model with a single parameter called the A distribution. The random variable X is said to follow a new distribution, with a single scale parameter β , trimmed as $X \sim A(\beta)$. The cumulative distribution function (CDF) is defined as follows:

$$F_X(x; \beta) = \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta}{x} \right) \right) \right), \quad x > 0, \beta > 0. \quad (4.3)$$

The cumulative distribution function (CDF) can be derived from the inverse Gompertz model as an exclusive scenario. The probability density function (PDF) of the A distribution can be expressed as

$$f_X(x; \beta) = \frac{1}{x^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta}{x} \right) \right) + \frac{\beta}{x} \right), \quad x > 0. \quad (4.4)$$

Putting $\alpha = 1$ in equations (4.1) and (4.2), the CDF and PDF of the A distribution with the scale parameter β can be derived as in equations (4.3) and (4.4). This implies that the A distribution serves as a specific case of the IG distribution. In many practical scenarios, classical distributions often fail to provide an adequate fit

to real-world data. As a result, numerous generators have been developed to create new models that better capture the complexities of empirical data, as evidenced by the studies referenced in [99], [100], [101], [102], [103], [104], [105], [106], [107], [97], [108].

In alignment with this research trajectory, we propose a novel unit distribution derived from the inverse Gumpertz distribution through an appropriate transformation of random variables. This new distribution aims to expand the existing repertoire of unit interval models, offering enhanced flexibility and applicability for modeling bounded data. Using the properties of the inverse Gumpertz distribution and employing a suitable transformation, we seek to address the potential limitations of existing models while providing a robust tool for statistical analysis in various applied fields. Moreover, additional considerations may be necessary beyond the data specified within the interval $[0,1]$ to ensure a comprehensive analysis and interpretation. In practice, unit data may also include boundary cases, such as intervals $(0, 1]$ or $[0, 1]$, where data points may consist of many zeros, ones, or both. Such scenarios require specialized models, such as hurdle models, to account for excess boundary values. Several hurdle models have been proposed in the literature, including the generalized Poisson hurdle model [109], inflated beta distributions [110], inflated Kumaraswamy distributions [111], and the inflated unit-Birnbaum-Saunders distribution [112]. These models are particularly useful for analyzing datasets with a preponderance of boundary values, ensuring robust and accurate statistical inference.

4.3 Methodology

4.3.1 Unit Inverse Gompertz Distribution (UIGD)

The inverse Gompertz distribution (IGD) was originally defined in the interval $(0, \infty)$ with the cumulative distribution function (CDF):

$$F_X(x) = \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta}{x} \right) \right) \right), \quad x > 0, \quad \beta > 0.$$

To use this distribution in contexts where variables are restricted to $[0, 1]$, we can apply the following transformations:

$$Y = \frac{1}{1 + X}, \tag{4.5}$$

where $X \sim \text{IGD}$. This transformation ensures that as $X \rightarrow 0^+$, $Y \rightarrow 1$, and as $X \rightarrow \infty$, $Y \rightarrow 0$. Thus, Y is bounded between 0 and 1.

The CDF of Y , denoted as $F_Y(y)$, can be derived from the relationship between Y and X :

$$F_Y(y) = P(Y \leq y) = P \left(X \geq \frac{1-y}{y} \right)$$

Using the complement rule:

$$F_Y(y) = 1 - F_X \left(\frac{1-y}{y} \right)$$

Substituting the CDF of X into this expression gives:

$$F_Y(y) = 1 - \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) \right), \quad y \in (0, 1) \quad (4.6)$$

The PDF of Y , denoted as $f_Y(y)$, can be obtained using the relationship:

$$f_Y(y) = f_X \left(\frac{1-y}{y} \right) \cdot \left| \frac{d}{dy} \left(\frac{1-y}{y} \right) \right| \quad (4.7)$$

Derivative of $\frac{1-y}{y}$:

$$\frac{d}{dy} \left(\frac{1-y}{y} \right) = -\frac{1}{y^2}$$

Substituting the PDF of X , given by:

$$f_X(x) = \frac{1}{x^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta}{x} \right) \right) + \frac{\beta}{x} \right), \quad x > 0 \quad (4.8)$$

Substituting $x = \frac{1-y}{y}$ into $f_X(x)$ we get

$$f_Y(y) = \frac{1}{\left(\frac{1-y}{y} \right)^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) + \frac{\beta y}{1-y} \right) \cdot \frac{1}{y^2} \quad (4.9)$$

Simplify the expression:

$$\frac{1}{\left(\frac{1-y}{y} \right)^2} \cdot \frac{1}{y^2} = \frac{y^2}{(1-y)^2} \cdot \frac{1}{y^2} = \frac{1}{(1-y)^2}$$

Thus, the PDF of Y becomes:

$$f_Y(y) = \frac{1}{(1-y)^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) + \frac{\beta y}{1-y} \right), \quad y \in (0, 1) \quad (4.10)$$

The transformed Inverse Gompertz Distribution on $(0, 1)$ is defined by:

- CDF:

$$F_Y(y) = 1 - \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) \right), \quad y \in (0, 1)$$

- PDF:

$$f_Y(y) = \frac{1}{(1-y)^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) + \frac{\beta y}{1-y} \right), \quad y \in (0, 1)$$

These expressions are suitable for modeling probabilities or proportions in a GLM.

Beyond the previously introduced transformation, an alternative mapping of X that also ensures its support is confined to the unit interval $(0, 1)$ is given by

$$Y = \frac{X}{1+X}, \quad X \sim \text{IGD}.$$

This transformation provides a different perspective on rescaling X while preserving its key distributional properties. Unlike the earlier transformation, which maps large values of X to points closer to 0, this transformation ensures a direct proportionality between X and Y , meaning that larger values of X correspond to values of Y closer to 1. Specifically, this transformation ensures that as $X \rightarrow 0^+$, $Y \rightarrow 0$, and as $X \rightarrow \infty$, $Y \rightarrow 1$. This can be particularly useful in applications where the relative size of

X plays a role in determining probabilities within the unit interval. In addition, this transformation maintains interpretability in contexts where Y represents a normalized or fractional version of X , such as proportion-based models or probability-related frameworks. By leveraging this alternative transformation, we derive the cumulative distribution function (CDF) of Y , denoted as $F_Y(y)$, as follows:

$$F_Y(y) = P(Y \leq y) = P\left(X \leq \frac{y}{1-y}\right)$$

Using the CDF of X , we get:

$$F_Y(y) = F_X\left(\frac{y}{1-y}\right) \quad (4.11)$$

Substituting the CDF of X into this expression gives:

$$F_Y(y) = \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta y}{1-y}\right)\right)\right), \quad y \in (0, 1) \quad (4.12)$$

The PDF of Y , denoted as $f_Y(y)$, can be obtained using the relationship:

$$f_Y(y) = f_X\left(\frac{y}{1-y}\right) \cdot \left|\frac{d}{dy}\left(\frac{y}{1-y}\right)\right| \quad (4.13)$$

Derivative of $\frac{y}{1-y}$:

$$\frac{d}{dy} \left(\frac{y}{1-y} \right) = \frac{1}{(1-y)^2}$$

Substituting the PDF of X :

$$f_X(x) = \frac{1}{x^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta}{x} \right) \right) + \frac{\beta}{x} \right), \quad x > 0 \quad (4.14)$$

Substituting $x = \frac{y}{1-y}$ into $f_X(x)$, we get:

$$f_Y(y) = \frac{1}{\left(\frac{y}{1-y} \right)^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) + \frac{\beta(1-y)}{y} \right) \cdot \frac{1}{(1-y)^2} \quad (4.15)$$

$$f_Y(y) = \frac{(1-y)^2}{y^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) + \frac{\beta(1-y)}{y} \right) \cdot \frac{1}{(1-y)^2} \quad (4.16)$$

Thus, the PDF of Y becomes:

$$f_Y(y) = \frac{1}{y^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) + \frac{\beta(1-y)}{y} \right), \quad y \in (0, 1) \quad (4.17)$$

The transformed Inverse Gompertz Distribution on $[0, 1]$ is defined by:

- CDF:

$$F_Y(y) = \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) \right), \quad y \in (0, 1).$$

- PDF:

$$f_Y(y) = \frac{1}{y^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) + \frac{\beta(1-y)}{y} \right), \quad y \in (0, 1).$$

These expressions are suitable for modeling probabilities or proportions in a GLM.

The plot of CDFs and PDFs for these two transformation methods based on different values of β are given in Figure 4.1

4.3.2 Comparison of Transformations

Both transformations, $Y = \frac{1}{1+X}$ and $Y = \frac{X}{1+X}$, map the original domain of X (which is $[0, \infty)$ for the Inverse Gompertz Distribution) into the range $[0, 1]$. However, the two transformations differ in their behavior at the boundaries of the domain, their mathematical properties, and the situations in which they may be more appropriate.

The following is a detailed comparison of these two transformations.

Transformation 1: $Y = \frac{1}{1+X}$

This transformation maps the variable X from the range $[0, \infty)$ to $[0, 1]$ in a decreasing monotonic way. Specifically, as $X \rightarrow 0^+$, $Y \rightarrow 1$ and as $X \rightarrow \infty$, $Y \rightarrow 0$. The CDF of Y , $F_Y(y)$ and the PDF of Y , $f_Y(y)$, are derived as follows:

$$F_Y(y) = 1 - \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta y}{1-y}\right)\right)\right), \quad y \in (0, 1)$$

$$f_Y(y) = \frac{1}{(1-y)^2} \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta y}{1-y}\right)\right) + \frac{\beta y}{1-y}\right), \quad y \in (0, 1)$$

Transformation 2: $Y = \frac{X}{1+X}$

This transformation also maps the variable X to the range $[0, 1]$, but in a different way, as $X \rightarrow 0^+$, $Y \rightarrow 0$ and as $X \rightarrow \infty$, $Y \rightarrow 1$. The CDF and PDF of Y are derived as:

$$F_Y(y) = \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta(1-y)}{y}\right)\right)\right), \quad y \in (0, 1)$$

$$f_Y(y) = \frac{1}{y^2} \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta(1-y)}{y}\right)\right) + \frac{\beta(1-y)}{y}\right), \quad y \in (0, 1)$$

4.3.3 Differences Between the Two Transformations

- Behavior at Boundaries:

1. $Y = \frac{1}{1+X}$ maps $X = 0$ to $Y = 1$, which means that it starts with a value of 1 and decreases to 0 as $X \rightarrow \infty$. This transformation might be useful

when higher values of X (and hence higher values of Y) are more likely or significant in a given model.

2. $Y = \frac{X}{1+X}$ maps $X = 0$ to $Y = 0$, which means that it starts at zero and increases to 1 as $X \rightarrow \infty$. This might be more suitable when the lower values of X (and hence the lower values of Y) are more frequent in a modeling scenario.

- Monotonicity: Both transformations are monotonic, but the directions are different. The first transformation decreases, while the second transformation increases.

- Speed of change:

1. In $Y = \frac{1}{1+X}$, the change in Y is more rapid for smaller values of X and the lower values increase X . This makes it useful when the data are clustered around smaller values of X .

2. In $Y = \frac{X}{1+X}$, the change in Y is gradual for smaller values of X and accelerates as X increases. This might be appropriate when the data are spread more evenly over larger values of X .

- Interpretation and Modeling Considerations:

1. $Y = \frac{1}{1+X}$ may be more appropriate when modeling situations where the outcome variable (which is constrained between 0 and 1) is more likely to take higher values and decrease as input X increases.

2. $Y = \frac{X}{1+X}$ might be more suitable when the outcome is expected to grow with input X , but growth slows as X increases.

4.3.4 Suitability for GLMs

Both transformations are useful in Generalized Linear Models (GLMs) when modeling variables bounded between 0 and 1. However, the choice of transformation should also consider the nature of the outcome variable, specifically regarding the presence of excess zeros and ones:

- **Excess Zeros:** If the outcome is more likely to be 0 (i.e., there are excess zeros in the data), the transformation $Y = \frac{X}{1+X}$ may be more appropriate. This is because this transformation starts at 0 when $X = 0$, and the values of Y increase toward 1 as X increases. It is better suited for situations where the data have a lower probability of success (that is, Y values close to 0) and gradually change to higher values as X increases.
- **Excess Ones:** If the outcome is more likely to be 1 (i.e., there are excess ones in the data), the transformation $Y = \frac{1}{1+X}$ is more appropriate. This transformation starts at 1 when $X = 0$ and decreases to 0 as X increases. It is more suitable when the outcome variable takes higher values, and the values decrease as the predictor X increases.

Both transformations are mathematically valid and allow for data modeling in the interval $[0, 1]$. The choice between $Y = \frac{1}{1+X}$ and $Y = \frac{X}{1+X}$ should be guided by the underlying behavior of the data and the direction of the relationship between the

predictor and the outcome variable. Additionally, if there are excess zeros or excess ones in the data:

- Transformation $Y = \frac{X}{1+X}$ is more appropriate for excess zeros.
- Transformation $Y = \frac{1}{1+X}$ is more suitable for excess ones.

Analyzing the distribution of the response variable and understanding the data domain will help determine which transformation is more appropriate for a given application.

4.3.5 Reliability and Hazard Rate Function

1. Reliability Function

The reliability function, also known as the survival function, is a fundamental concept in lifetime analysis and reliability engineering, serving as a key tool for quantifying the probability that a functional device or system will operate without failure for a specified period. It is widely used to evaluate the performance, durability, and lifespan of components, systems, and products. For a random variable Y representing the lifetime or time to failure of a component, the reliability function $R_Y(y)$ is defined as;

$$R_Y(y) = 1 - F_Y(y) \tag{4.18}$$

$F_Y(y)$ is the cumulative distribution function (CDF) of the components.

The function $F_Y(y)$, often termed unreliability, represents the probability of failure at time t . This concept forms the foundation of reliability engineering and survival analysis, enabling the assessment and prediction of system performance and failure over time [113]. The reliability function of the unit inverse Gumpertz distribution (UIGD) is expressed as;

From (4.18), using the first transformation, where $Y = \frac{1}{X+1}$

$$R_Y(y) = 1 - \left(1 - \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) \right) \right)$$

$$R_Y(y) = \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) \right) \quad (4.19)$$

Using the second transformation, where $Y = \frac{X}{X+1}$

$$R_Y(y) = 1 - \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) \right) \quad (4.20)$$

2. Hazard Rate Function (HRF)

The Hazard Rate Function (HRF), the reverse of the Reliability Function, also known as the instantaneous failure rate, is another critical measure in reliability analysis. Quantifies the likelihood of failure for a device or system that is currently operating. A higher hazard rate indicates a higher probability of

imminent failure, reflecting an increased risk at a specific point in time. Mathematically, the hazard function represents the probability of failure over a short interval t_0 to $t_0 + \delta t_0$, given that the system has survived up to t_0 . The hazard rate function can be written in terms of the probability density function (PDF) $f_Y(y)$ and the reliability function $R_Y(y)$:

$$h_Y(y) = \frac{f_Y(y)}{R_Y(y)} \quad (4.21)$$

The hazard function is particularly valuable for identifying periods of increased risk in a system's lifecycle, facilitating proactive maintenance and risk mitigation strategies. It is widely applied in engineering, healthcare, and manufacturing to predict failure behavior and enhance system reliability [55]. From (4.21), the hazard rate based on the first transformation is expressed as;

$$h_Y(y) = \frac{\frac{1}{(1-y)^2} \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta y}{1-y}\right)\right) + \frac{\beta y}{1-y}\right)}{\exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta y}{1-y}\right)\right)\right)}$$

$$h_Y(y) = \frac{1}{(1-y)^2} \exp\left(\frac{\beta y}{1-y}\right) \quad (4.22)$$

For the second transformation, the hazard rate

$$h_Y(y) = \frac{\frac{1}{y^2} \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta(1-y)}{y}\right)\right) + \frac{\beta(1-y)}{y}\right)}{1 - \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta(1-y)}{y}\right)\right)\right)} \quad (4.23)$$

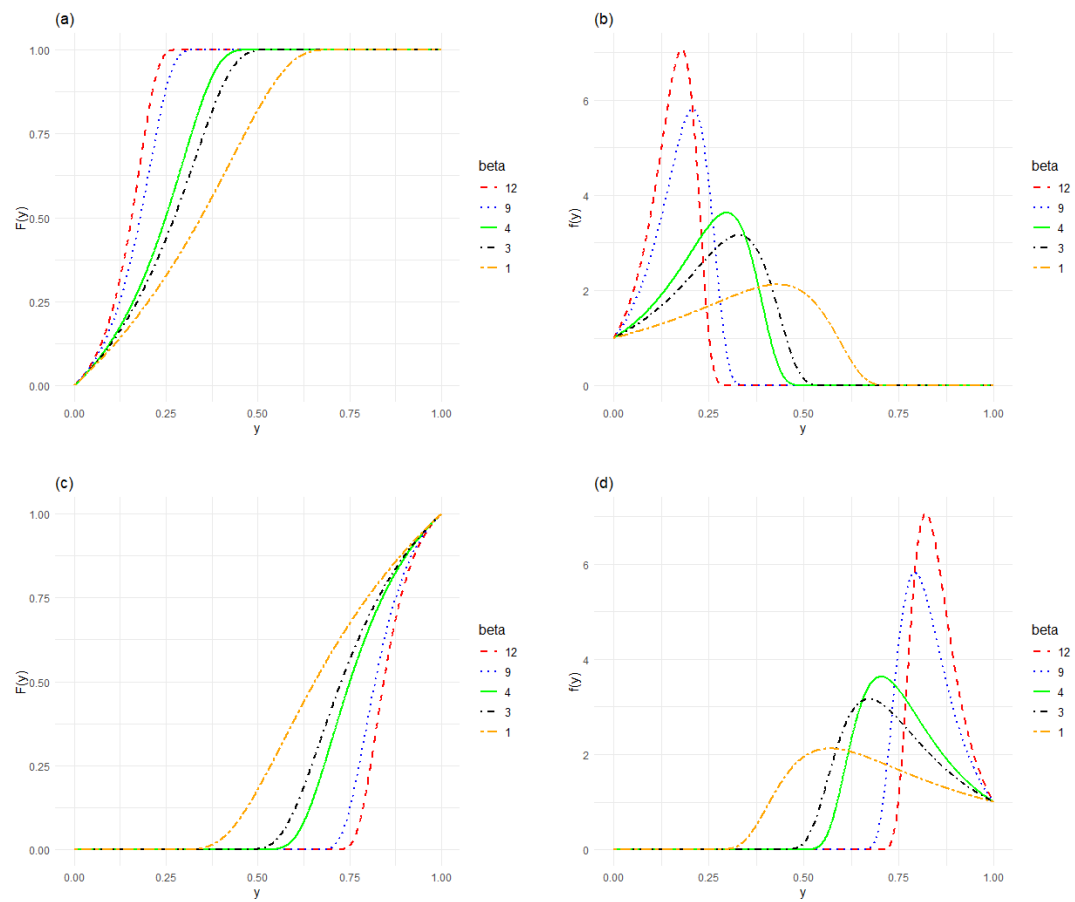


Figure 4.1: Some plots of the cdf (a) and (c), pdf(b) and (d) of the proposed distribution for selected values of β of the two transformations

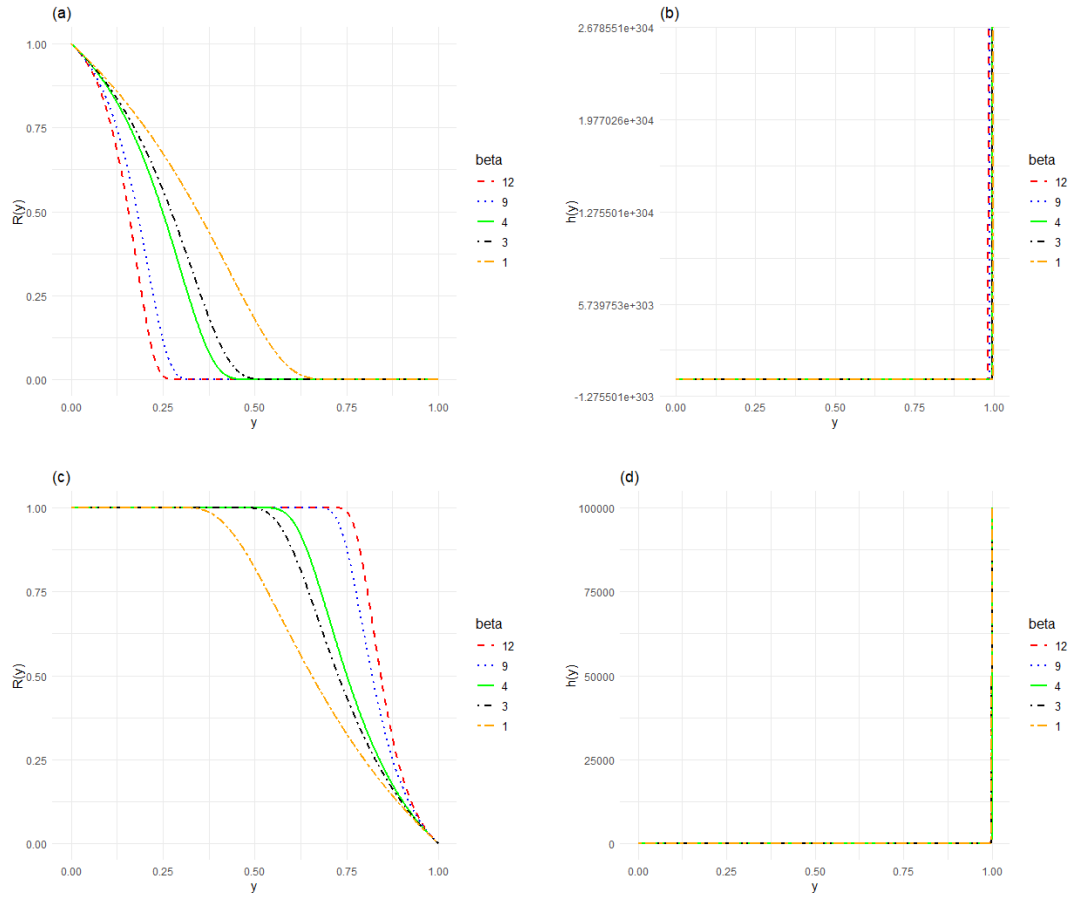


Figure 4.2: Some plots of the reliability function (a) and (c), and hazard rate function (b) and (d) of the proposed distribution for selected values of β of the two transformations

4.3.6 The p -th Ordinary Moment

The p -th ordinary moment of a continuous random variable Y is defined as:

$$\mathbb{E}(Y^p) = \int_{y \in \mathbb{R}} y^p f(y) dy, \quad (4.24)$$

where $f(y)$ is the probability density function (PDF) of Y . If Y follows the unit inverse Gompertz distribution under the first transformation:

$$Y = \frac{1}{1 + X},$$

Where X is a random variable with the inverse Gompertz distribution, then the p -th moment of the new unit distribution is given by:

$$E[Y^p] = \int_0^1 y^p \frac{1}{(1-y)^2} \exp\left(\frac{1}{\beta} \left(1 - \exp\left(\frac{\beta y}{1-y}\right)\right) + \frac{\beta y}{1-y}\right) dy \quad (4.25)$$

Using the Taylor series expansion of the exponential function:

$$\exp\left(\frac{\beta y}{1-y}\right) = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{\beta y}{1-y}\right)^n.$$

we rewrite:

$$1 - \exp\left(\frac{\beta y}{1-y}\right) = - \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\beta y}{1-y}\right)^n. \quad (4.26)$$

Now, if we multiplying equation (4.26) by $\frac{1}{\beta}$ we have:

$$\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) = \frac{1}{\beta} - \frac{1}{\beta} \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\beta y}{1-y} \right)^n. \quad (4.27)$$

Adding the term $\frac{\beta y}{1-y}$ in equation (4.27), we have:

$$\frac{1}{\beta} - \frac{1}{\beta} \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\beta y}{1-y} \right)^n + \frac{\beta y}{1-y}. \quad (4.28)$$

Exponentiating (4.28), we have:

$$\exp \left(\frac{1}{\beta} - \frac{1}{\beta} \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\beta y}{1-y} \right)^n + \frac{\beta y}{1-y} \right). \quad (4.29)$$

Using the Taylor expansion:

$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!},$$

equation (4.29) becomes:

$$\sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{1}{\beta} - \frac{1}{\beta} \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\beta y}{1-y} \right)^n + \frac{\beta y}{1-y} \right)^m. \quad (4.30)$$

Substituting (4.30) back to (4.25) we have:

$$E[Y^p] = \int_0^1 y^p \frac{1}{(1-y)^2} \sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{1}{\beta} - \frac{1}{\beta} \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\beta y}{1-y} \right)^n + \frac{\beta y}{1-y} \right)^m dy. \quad (4.31)$$

Now, it is not difficult to see that the integral in (4.31) can be rewritten as:

$$E[Y^p] = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} C_{m,n} \boxed{\int_0^1 y^{p+n} (1-y)^{-2-n} dy}. \quad (4.32)$$

Recognizing that the boxed part of (4.32) is a Beta function and can be expressed as,

$$\int_0^1 y^{a-1} (1-y)^{b-1} dy = B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

with $a = p + n + 1$ and $b = -1 - n$, we obtain:

$$B(p + n + 1, -1 - n) = \frac{\Gamma(p + n + 1)\Gamma(-1 - n)}{\Gamma(p + 1)}.$$

Thus,

$$E[Y^p] = \sum_{m=0}^{\infty} C_m \frac{\Gamma(p + 1)\Gamma(m + 1)}{\Gamma(p + m + 2)}. \quad (4.33)$$

The sum in equation (4.33) can be written in terms of a hypergeometric function,

hence the p th moment based on the transformation $Y = \frac{1}{1+X}$ becomes

$$E[Y^p] = \frac{\Gamma(p + 1)}{\Gamma(2)} {}_2F_1(1, p + 1; p + 2; -\beta). \quad (4.34)$$

Similarly, if Y follows the unit inverse Gompertz distribution under the second transformation, it is defined as:

$$Y = \frac{X}{1 + X},$$

$$E(Y^p) = \int_0^1 y^p \frac{1}{y^2} \left[\exp \left[\beta \left(\frac{1}{y} - 1 \right) \right] \cdot \exp \left[-\frac{e^{\beta(\frac{1}{y}-1)} - 1}{\beta} \right] \right] dy$$

Let:

$$t = e^{\beta(\frac{1}{y}-1)} \implies \frac{dt}{dy} = -\frac{\beta e^{\beta(\frac{1}{y}-1)}}{y^2} \implies -\frac{dt}{\beta} = \frac{e^{\beta(\frac{1}{y}-1)}}{y^2}$$

Also:

$$\ln t = \beta \left(\frac{1}{y} - 1 \right) \implies \frac{\beta}{\ln t + \beta} = y$$

Then:

$$E(Y^p) = - \int_{\infty}^1 \left(\frac{\beta}{\ln t + \beta} \right)^p e^{-\frac{t-1}{\beta}} \times \frac{dt}{\beta}$$

$$= \beta^{p-1} e^{\frac{1}{\beta}} \left| \int_1^{\infty} \left(\frac{1}{\ln t + \beta} \right)^p e^{-\frac{t}{\beta}} dt \right|$$

The given integral is:

$$I = \int_1^{\infty} \left(\frac{1}{\ln t + \beta} \right)^p e^{-\frac{t}{\beta}} dt.$$

The exponential term $e^{-\frac{t}{\beta}}$ can be expanded as a Taylor series:

$$e^{-\frac{t}{\beta}} = \sum_{k=0}^{\infty} \frac{\left(-\frac{t}{\beta}\right)^k}{k!} = \sum_{k=0}^{\infty} \frac{(-1)^k t^k}{\beta^k k!}.$$

Substituting this into the integral:

$$I = \int_1^{\infty} \left(\frac{1}{\ln t + \beta} \right)^p \sum_{k=0}^{\infty} \frac{(-1)^k t^k}{\beta^k k!} dt.$$

Interchanging the summation and integrating (valid under uniform convergence):

$$I = \sum_{k=0}^{\infty} \frac{(-1)^k}{\beta^k k!} \int_1^{\infty} \frac{t^k}{(\ln t + \beta)^p} dt.$$

Let:

$$I_k = \int_1^{\infty} \frac{t^k}{(\ln t + \beta)^p} dt.$$

To simplify I_k , we substitute $\ln t = u$, so that $t = e^u$ and $dt = e^u du$. The limits change as follows:

$$t = 1 \implies u = 0, \quad t \rightarrow \infty \implies u \rightarrow \infty.$$

This transforms I_k into:

$$I_k = \int_0^\infty \frac{e^{ku}}{(u+\beta)^p} e^u du = \int_0^\infty \frac{e^{(k+1)u}}{(u+\beta)^p} du.$$

We can expand $\frac{1}{(u+\beta)^p}$ by using the binomial series expansion for negative powers:

$$\frac{1}{(u+\beta)^p} = \frac{1}{\beta^p} \left(1 + \frac{u}{\beta}\right)^{-p} = \frac{1}{\beta^p} \sum_{m=0}^{\infty} \binom{-p}{m} \left(\frac{u}{\beta}\right)^m.$$

Substituting this into I_k :

$$I_k = \frac{1}{\beta^p} \int_0^\infty e^{(k+1)u} \sum_{m=0}^{\infty} \binom{-p}{m} \left(\frac{u}{\beta}\right)^m du.$$

$$I_k = \frac{1}{\beta^p} \sum_{m=0}^{\infty} \binom{-p}{m} \frac{1}{\beta^m} \int_0^\infty u^m e^{(k+1)u} du.$$

The integral $\int_0^\infty u^m e^{(k+1)u} du$ is a standard form of the Gamma function after substituting $v = -(k+1)u$:

$$\int_0^\infty u^m e^{(k+1)u} du = \frac{(-1)^m m!}{(k+1)^{m+1}}.$$

Substituting into I_k :

$$I_k = \frac{1}{\beta^p} \sum_{m=0}^{\infty} \binom{-p}{m} \frac{m!}{\beta^m} \frac{(-1)^m}{(k+1)^{m+1}} \quad (4.35)$$

Substituting I_k into I , we have:

$$I = \sum_{k=0}^{\infty} \frac{(-1)^k}{\beta^k k!} \frac{1}{\beta^p} \sum_{m=0}^{\infty} \binom{-p}{m} \frac{m!}{\beta^m} \frac{(-1)^m}{(k+1)^{m+1}}.$$

$$E(Y^p) = \beta^{p-1} e^{\frac{1}{\beta}} \times I$$

$$= \beta^{p-1} e^{\frac{1}{\beta}} \sum_{k=0}^{\infty} \frac{(-1)^k}{\beta^k k!} \frac{1}{\beta^p} \sum_{m=0}^{\infty} \binom{-p}{m} \frac{m!}{\beta^m} \frac{(-1)^m}{(k+1)^{m+1}}.$$

The hypergeometric function ${}_2F_1(a, b; c; z)$ is defined as:

$${}_2F_1(a, b; c; z) = \sum_{m=0}^{\infty} \frac{(a)_m (b)_m}{(c)_m} \frac{z^m}{m!}, \quad (4.36)$$

Where $(a)_m$ is the Pochhammer symbol, defined as:

$$(a)_m = a(a+1)(a+2) \cdots (a+m-1) = \frac{\Gamma(a+m)}{\Gamma(a)}.$$

The double summation in the expression is the following:

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{\beta^k k!} \sum_{m=0}^{\infty} \binom{-p}{m} \frac{m!}{\beta^m} \frac{(-1)^m}{(k+1)^{m+1}}. \quad (4.37)$$

The term $\binom{-p}{m}$ can be written in terms of the Pochhammer symbol:

$$\binom{-p}{m} = \frac{(-p)(-p-1) \cdots (-p-m+1)}{m!} = \frac{(-1)^m (p)_m}{m!}.$$

Substituting this into the summation gives:

$$\sum_{m=0}^{\infty} \binom{-p}{m} \frac{m!}{\beta^m} \frac{(-1)^m}{(k+1)^{m+1}} = \sum_{m=0}^{\infty} \frac{(-1)^m (p)_m}{\beta^m m!} \frac{m!}{(k+1)^{m+1}}.$$

Simplifying the factorials yields the following:

$$\sum_{m=0}^{\infty} \frac{(p)_m}{(k+1)^{m+1} \beta^m}.$$

Factoring in $\frac{1}{k+1}$:

$$\frac{1}{k+1} \sum_{m=0}^{\infty} \frac{(p)_m}{(k+1)^m \beta^m}.$$

$${}_2F_1 \left(p, 1; 1; \frac{1}{\beta(k+1)} \right).$$

$$E(Y^p) = \beta^{p-1} e^{\frac{1}{\beta}} \sum_{k=0}^{\infty} \frac{(-1)^k}{\beta^k k!} \frac{1}{\beta^p (k+1)} {}_2F_1 \left(p, 1; 1; \frac{1}{\beta(k+1)} \right). \quad (4.38)$$

The p -th moment of Y is given by:

$$E(Y^p) = \beta^{p-1} e^{\frac{1}{\beta}} \sum_{k=0}^{\infty} \frac{(-1)^k}{\beta^k k!} \frac{1}{\beta^p (k+1)} {}_2F_1 \left(p, 1; 1; \frac{1}{\beta(k+1)} \right), \quad (4.39)$$

where ${}_2F_1(a, b; c; z)$ is the hypergeometric function.

4.3.7 Parameter Estimation

Several methods are available to estimate the unknown parameters of a given statistical model. Some of the most widely used methods in the literature include the method of moments (MoM), the L-moments approach, maximum likelihood estimation (MLE), the least squares approach, probability-weighted moments (PWM), and the mixed moments approach. For further details on these methodologies, see [114],[115], [116], and [117]. Among these, Maximum Likelihood Estimation (MLE) is regarded as one of the most effective techniques for parameter estimation due to the following advantages: Mathematical Simplicity, Asymptotic Unbiasedness, Minimal Variance, and Invariance Property. The MLE of a transformed function or parameter is equivalent to the transformation of the MLE of the original function or parameter. For additional information on MLE, see [118],[119], and [120].

4.3.8 Maximum Likelihood Estimation for UIG Distribution

This research used the maximum likelihood estimate (MLE) method to estimate the proposed unit-inverse Gompertz distribution parameter. Assuming Y_i to be a random variable with the unit inverse Gompertz distribution, the likelihood function is given as:

$$L(\Theta) = \prod_{i=1}^n f(y_i; \beta) \quad (4.40)$$

Substituting Equation (4.17) into Equation (4.40), we have the following.

$$L(\Theta) = \prod_{i=1}^n \frac{1}{y_i^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y_i)}{y_i} \right) \right) + \frac{\beta(1-y_i)}{y_i} \right) \quad (4.41)$$

In practice, it is easier to work with the log-likelihood function. The log-likelihood function $\ell(\Theta|\mathbf{y})$ based on a random sample of size n with the corresponding observations $y_1, y_2, \dots, y_n$ are given by:

$$\ell(\Theta|\mathbf{y}) = \log L(\Theta) = \sum_{i=1}^n \log f(y_i; \beta) \quad (4.42)$$

We obtain the following by substituting Equation (4.17) into Equation (4.42).

$$\ell(\Theta|\mathbf{y}) = \sum_{i=1}^n \left[-\log(y_i^2) + \frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y_i)}{y_i} \right) \right) + \frac{\beta(1-y_i)}{y_i} \right] \quad (4.43)$$

Simplifying Equation (4.43), we have:

$$\ell(\Theta|\mathbf{y}) = \sum_{i=1}^n \left[-2\log(y_i) + \frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y_i)}{y_i} \right) \right) + \frac{\beta(1-y_i)}{y_i} \right] \quad (4.44)$$

The partial derivative of $\ell(\Theta|\mathbf{y})$ given in Equation (4.44) with respect to β is:

$$\frac{\partial \ell(\Theta|\mathbf{y})}{\partial \beta} = \sum_{i=1}^n \left[\frac{-1}{\beta^2} \left(1 - \exp \left(\frac{\beta(1-y_i)}{y_i} \right) \right) - \frac{1}{\beta} \left(\exp \left(\frac{\beta(1-y_i)}{y_i} \right) \right) \frac{(1-y_i)}{y_i} + \frac{(1-y_i)}{y_i} \right] \quad (4.45)$$

Similarly, we obtain the following by substituting Equation (4.17) into Equation (2.36).

$$L(\Theta) = \prod_{i=1}^n \frac{1}{(1-y_i)^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y_i}{1-y_i} \right) \right) + \frac{\beta y_i}{1-y_i} \right) \quad (4.46)$$

The log-likelihood function is then given by:

$$\ell(\Theta|\mathbf{y}) = \sum_{i=1}^n \left[-\log(1-y_i)^2 + \frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y_i}{1-y_i} \right) \right) + \frac{\beta y_i}{1-y_i} \right] \quad (4.47)$$

Simplifying Equation (4.47), we obtain:

$$\ell(\Theta|\mathbf{y}) = \sum_{i=1}^n \left[-2\log(1-y_i) + \frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y_i}{1-y_i} \right) \right) + \frac{\beta y_i}{1-y_i} \right] \quad (4.48)$$

The partial derivative of $\ell(\Theta|\mathbf{y})$ in Equation (4.48) with respect to β is given by:

$$\frac{\partial \ell(\Theta|\mathbf{y})}{\partial \beta} = \sum_{i=1}^n \left[\frac{-1}{\beta^2} \left(1 - \exp \left(\frac{\beta y_i}{1-y_i} \right) \right) - \frac{1}{\beta} \exp \left(\frac{\beta y_i}{1-y_i} \right) \left(\frac{y_i}{1-y_i} \right) + \frac{y_i}{1-y_i} \right] \quad (4.49)$$

It is important to note that Equations (4.45) and (4.49) are nonlinear for the parameter β . Consequently, an analytical solution cannot be derived by setting the equation to zero. However, maximum likelihood estimates (MLE) for β , denoted as $\hat{\beta}$, can be efficiently computed using numerical optimization methods. One such approach is the Newton-Raphson algorithm, which minimizes the log-likelihood function $\ell(\Theta|\mathbf{y})$. Various statistical and mathematical software packages, such as the `nlm` and `optim` functions in R [121], provide robust implementations of Newton-type algorithms for this purpose.

4.3.9 Model Adequacy

To evaluate the suitability of a parametric model fitted to the data, it is essential to determine how well it conforms to the observed data. One widely used approach for this purpose is the Kolmogorov-Smirnov (KS) test, which is formulated as:

$$KS = \max_{y \in \text{Data}} \left| F_n(y) - \hat{F}(y) \right| \quad (4.50)$$

This test helps to assess the goodness of fit of continuous univariate distributions, particularly those with a support range of $\mathbb{R}^+$. An extensive value of the KS test p (greater than 0.05) implies that the assumed distribution adequately represents the data. Additionally, the adequacy of nonnested models can be compared using information criteria metrics, such as:

- Akaike Information Criterion (AIC) [122]:

$$AIC = -2\mathcal{L} + 2k$$

- Bayesian Information Criterion (BIC) [123]:

$$BIC = -2\mathcal{L} + k \log(n)$$

- Akaike Information Criterion with Correction (AICc) [124]:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

where $\mathcal{L}$ denotes the maximized log-likelihood value, k represents the number of model parameters, and n is the sample size. A model with the lowest AIC, BIC, and AICc values is generally preferred, indicating better model performance than others.

4.3.10 Unit Inverse Gompertz Hurdle distribution (UIGHurdle)

The unit inverse Gompertz distribution is well-suited for modeling random variables defined on the open interval $[0, 1)$, especially when the data contain few or no zeros. However, in practical applications, we often encounter datasets that include a significant number of zeros and ones or even an excess of these values. In such scenarios, the support of the random variable extends to the closed interval $[0, 1]$, necessitating a

model capable of accommodating the presence of these boundary values. Specifically, a hurdle model is required to effectively capture the behavior of zeros and ones in the data. Notably, in the literature, many continuous distributions are defined with probability density functions (pdfs) that exclude the endpoints of their support (e.g., 0 and 1), even though these distributions may theoretically assign non-zero densities to these points. This exclusion can lead to limitations in modeling real-world data, where zeros and ones are often observed. Therefore, developing or adapting models that explicitly account for these boundary values is essential for accurate and realistic data analysis.

A few examples include the following: the exponential distribution with pdf defined as $f(y) = \lambda e^{-\lambda y}$, $y, \lambda > 0$ in both [125] and [126] omits 0 at the support of y even when $f(0) = \lambda$; the Marshall–Olkin generalized exponential distribution with pdf defined as $f(y) = p\theta e^{-\theta y} [1 - (1-p)e^{-\theta y}]^{p-1}$, $y > 0$ and $p \in [0, 1]$ in [127] omits 0 at the support of y even when $f(0) = p\theta$; the unit Lindley distribution with pdf defined as $f(y) = \theta^2 / [1 + \theta](1-y)^2 \exp(-\theta y / [1-y])$, $\theta > 0$, $y \in (0, 1]$ in [72] omits 0 at the support of y even when it is clear that $f(0) = \theta^2 / (1 + \theta)$; the unit-improved second-degree Lindley distribution with pdf defined as $f(y) = \lambda z^3 / (\lambda^2 + 2\lambda + 2)(1 + y/(1 - y))^2 \exp(-\lambda y / (1 - y))$, $\lambda > 0$, $y \in (0, 1]$ in [138] omits 0 at the support of y even when it is obvious that $f(0) = z^3 / (\lambda^2 + 2\lambda + 2)$; and the unit-Gompertz distribution with pdf defined as $f(y) = \alpha\beta y^{\beta-1} \exp(-\alpha y^\beta)$, $\alpha, \beta > 0$, and $y \in (0, 1]$ in [62] omits 1 at the support of y even when $f(1) \neq e^\alpha$. The beta distribution, a widely used model for random variables defined on the unit interval, has two main variants in terms of its support. Some authors define the beta distribution with support on the closed interval ($y \in$

$[0, 1]$), explicitly including the endpoints, while others restrict the support to the open interval ($y \in (0, 1)$), excluding the endpoints. This distinction is important because it influences how the distribution is applied in practice, particularly when modeling data that include values at the boundaries (i.e., 0 and 1). For example, as discussed in <https://www.sciencedirect.com/topics/mathematics/beta-distribution>, the choice of support can affect the interpretation and applicability of the beta distribution in various contexts.

In cases where the data exhibit inflation at 0 and 1 (i.e., an excess of observations at these points), the standard beta distribution with support on ($y \in [0, 1]$) may not be sufficient. The inflated beta distribution was introduced by [110] to address this limitation. This model extends the traditional beta distribution by incorporating mechanisms to account for the inflated probabilities at 0 and 1, making it more suitable for datasets with significant boundary values. By combining the continuous beta distribution on ($y \in (0, 1)$) with discrete components at 0 and 1, the inflated beta distribution provides a flexible framework for modeling data that spans the entire closed interval ($y \in [0, 1]$). This approach highlights the importance of carefully considering the support of a distribution and adapting models to better reflect real-world data's characteristics. Hence, the random variable Y could be said to follow the unit inverse Gompertz distribution of the first transformation if its CDF is defined as:

$$F_Y(y) = 1 - \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1 - y} \right) \right) \right), \quad y \in (0, 1), \quad (4.51)$$

And PDF is defined as

$$f_Y(y) = \frac{1}{(1-y)^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y}{1-y} \right) \right) + \frac{\beta y}{1-y} \right), \quad y \in (0, 1). \quad (4.52)$$

Second transformation,

$$F_Y(y) = \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) \right), \quad y \in (0, 1), \quad (4.53)$$

$$f_Y(y) = \frac{1}{y^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1-y)}{y} \right) \right) + \frac{\beta(1-y)}{y} \right), \quad y \in (0, 1). \quad (4.54)$$

From this point onward, Equations (4.51), (4.52), (4.53), and (4.54) will be referred to as the cumulative distribution function (CDF) and probability density function (PDF) of the unit inverse Gompertz distribution. However, when the random variable Y generates a significant number of 0's and 1's in addition to values within the interval $(0, 1)$, the standard Unit Inverse Gompertz distribution is inadequate for modeling such data. This is because the Unit Inverse Gompertz distribution, as defined in Equations (4.52) and (4.54), does not account for the presence of boundary values (0 and 1). To address this limitation, we propose an extension of the Unit Inverse Gompertz distribution. This extended model incorporates mechanisms to handle the inflation of 0s and 1s, making it suitable for datasets that include a substantial propor-

tion of these boundary values. By combining the continuous Unit Inverse Gompertz distribution for values within $(0, 1)$ with discrete components for 0 and 1, the extended model provides a comprehensive framework for modeling data that spans the entire closed interval $[0, 1]$. This adaptation ensures that the model can accurately capture the behavior of datasets with an excess of 0s and 1s, offering a more robust and flexible approach for practical applications. The UIGHurdle distribution is a mixture model whose cdf is defined by;

$$F_{\text{UIGHurdle}}(y; \phi, \psi, \beta) = \phi \text{Ber}(y; \psi) + (1 - \phi)F(y; \beta),$$

and PDF defined by

$$f_{\text{UIGHurdle}}(y; \phi, \psi, \beta) = \begin{cases} \phi(1 - \psi), & \text{if } y = 0, \\ (1 - \phi)f(y; \beta), & \text{if } y \in (0, 1), \\ \phi\psi, & \text{if } y = 1, \end{cases} \quad (4.55)$$

where β is as defined before, $y \in [0, 1]$, $\psi \in [0, 1]$, $\phi \in (0, 1)$ is the proportion of degenerate values that are equal to 0 or 1, ϕ is the mixing parameter (i.e., the proportion of y that are either 0 or 1), $\text{Ber}(y; \psi)$ denotes the cdf of a Bernoulli random variable with a parameter ψ , $F(\cdot)$ and $f(\cdot)$ denote the cdf and pdf of the unit inverse gompertz distribution in Equation (4.51), (4.52), and Equation (4.53), (4.54) respec-

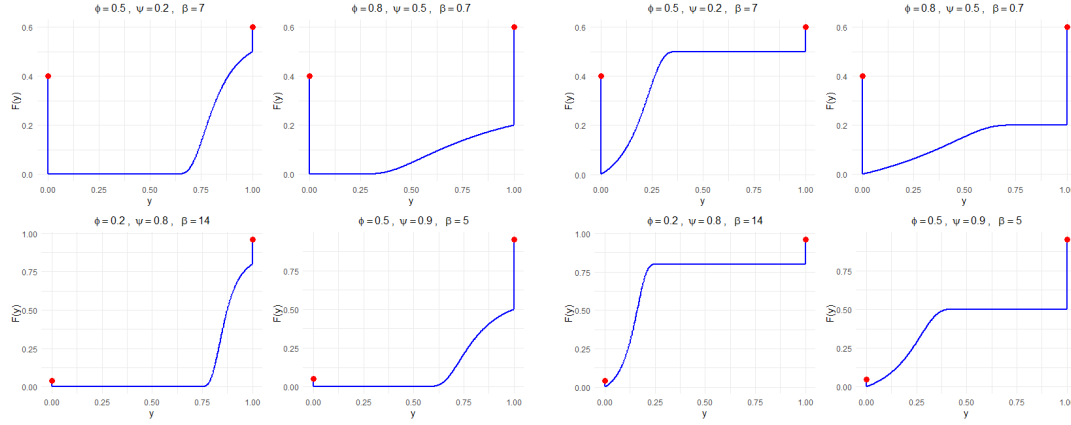


Figure 4.3: Some plots of the CDF of the proposed UIGHurdle distribution for selected values of ϕ, ψ , and β of the two transformations

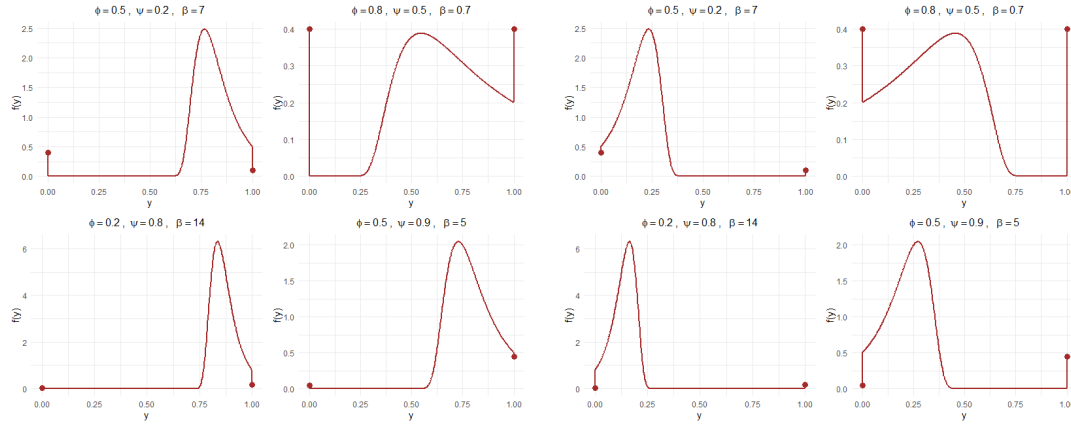


Figure 4.4: Some plots of the PDF of the proposed UIGHurdle distribution for selected values of ϕ, ψ , and β of the two transformations

tively. Figures 10 and 11 show the cdf and pdf plots of the UIGHurdle distribution for different combinations of parameters.

If Y follows the UIGhurdle distribution, its k th order ordinary moment is given by;

$$E(Y^k) = \phi\psi + (1 - \phi)\mu'_k,$$

where μ'_k is the k th ordinary moment of the unit inverse Gompertz distribution.

In particular, the probability density function (pdf) of the UIGhurdle distribution in Equation (4.55) can be expressed in a more compact or interpretable form. This alternative representation often provides deeper insights into the distribution structure and its components, and can be described as

$$\begin{aligned} f_{\text{UIGHurdle}}(y; \phi, \psi, \beta) &= [\phi\psi(1 - \psi)^{1-I_{(0,1)}(y)}] \times [(1 - \phi)f(y; \beta)]^{1-I_{(0,1)}(y)} \\ &= [\phi^{I_{(0,1)}(y)}(1 - \phi)^{1-I_{(0,1)}(y)}] \times [\psi^y(1 - \psi)^{1-y}]^{1-I_{(0,1)}(y)} \times f(y; \beta)^{1-I_{(0,1)}(y)}, \quad (4.56) \end{aligned}$$

Here, $I_{(0,1)}(y)$ is an indicator function that takes the value of 1 if $y \in (0, 1)$ and 0 otherwise.

4.3.11 Unit UTW Distribution

The unit upper truncated Weibull distribution was also incorporated into this research to effectively capture the data's characteristics. This distribution, known for

its adaptability to model various risk rates, was particularly useful for analyzing data sets where values are constrained within the interval $[0, 1]$. The CDF of the UUTW distribution is defined by

$$F(y) = \frac{1 - \exp(-\lambda y^\beta)}{1 - \exp(-\lambda)}$$

With the corresponding PDF given by

$$f(y) = \frac{\lambda\beta}{1 - \exp(-\lambda)} y^{\beta-1} \exp(-\lambda y^\beta)$$

To refine the analytical framework, the unit upper truncated Weibull distribution was transformed into a hurdle model, which is particularly effective in handling datasets with an excess of zeros and ones. The UUTWHurdle distribution is a mixture model whose cdf is defined by;

$$F_{\text{UUTWHurdle}}(y; \phi, \psi, \beta, \lambda) = \phi \text{Ber}(y; \psi) + (1 - \phi)F(y; \lambda, \beta),$$

and PDF defined by

$$f_{\text{UUTWHurdle}}(y; \phi, \psi, \beta, \lambda) = \begin{cases} \phi(1 - \psi), & \text{if } y = 0, \\ (1 - \phi)f(y; \lambda, \beta), & \text{if } y \in (0, 1), \\ \phi\psi, & \text{if } y = 1, \end{cases} \quad (4.57)$$

4.3.12 The MLE for the UIGHurdle Distribution

Suppose we draw the random sample $y = (y_1, y_2, \dots, y_n)'$ from the UIGhurdle distribution, the likelihood function for $\Phi = (\psi, \phi, \beta)'$ is given by

$$L(\Phi) = \prod_{i=1}^n f_{\text{UIGHurdle}}(y; \phi, \psi, \beta) = L_1(\phi|y) \times L_2(\psi|y) \times L_3(\beta|y), \quad (4.58)$$

where

$$L_1(\phi|y) = \prod_{i=1}^n \phi^{I_{(0,1)}(y_i)} (1 - \phi)^{1 - I_{(0,1)}(y_i)} = \phi^{\sum_{i=1}^n I_{(0,1)}(y_i)} (1 - \phi)^{n - \sum_{i=1}^n I_{(0,1)}(y_i)},$$

$$L_2(\psi|y) = \prod_{i=1}^n \psi^{y_i} (1 - \psi)^{1 - y_i} I_{(0,1)}(y_i) = \psi^{\sum_{i=1}^n y_i I_{(0,1)}(y_i)} (1 - \psi)^{\sum_{i=1}^n (1 - y_i) I_{(0,1)}(y_i)},$$

and

$$L_3a(\beta|y) = \prod_{i=1}^n f(y_i; \beta) = \prod_{i=1}^n \frac{1}{(1 - y_i)^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y_i}{1 - y_i} \right) \right) + \frac{\beta y_i}{1 - y_i} \right)$$

$$L_3b(\beta|y) = \prod_{i=1}^n f(y_i; \beta) = \prod_{i=1}^n \frac{1}{y_i^2} \exp \left(\frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1 - y_i)}{y_i} \right) \right) + \frac{\beta(1 - y_i)}{y_i} \right)$$

The log-likelihood function can be expressed as the natural logarithm of Equation (4.58) as in Equation (4.59):

$$\mathcal{L}(\Phi|y) = \ln \prod_{i=1}^n f_{\text{UIGhurdle}}(y; \phi, \psi, \beta) = \mathcal{L}_1(\phi|y) + \mathcal{L}_2(\psi|y) + \mathcal{L}_3(\beta|y), \quad (4.59)$$

where

$$\mathcal{L}_1(\phi|y) = \ln(\phi) \sum_{i=1}^n I_{(0,1)}(y_i) + \ln(1 - \phi) \left[n - \sum_{i=1}^n I_{(0,1)}(y_i) \right],$$

$$\mathcal{L}_2(\psi|y) = \ln(\psi) \sum_{i=1}^n y_i I_{(0,1)}(y_i) + \ln(1 - \psi) \sum_{i=1}^n (1 - y_i) I_{(0,1)}(y_i),$$

and

$$\ell_3 a(\beta|y) = \sum_{i=1}^n \left[-2 \log(1 - y_i) + \frac{1}{\beta} \left(1 - \exp \left(\frac{\beta y_i}{1 - y_i} \right) \right) + \frac{\beta y_i}{1 - y_i} \right]$$

$$\ell_3 b(\beta|y) = \sum_{i=1}^n \left[-2 \log(y_i) + \frac{1}{\beta} \left(1 - \exp \left(\frac{\beta(1 - y_i)}{y_i} \right) \right) + \frac{\beta(1 - y_i)}{y_i} \right]$$

We conclude this section by emphasizing that, with slight modifications to Equation (4.56), the proposed model can be adapted to handle cases that involve only a zero hurdle (that is, $y \in [0, 1)$) or only one hurdle (that is, $y \in (0, 1]$). Specifically, by adjusting the probability mass assignments at the boundaries, the model can be tailored

to address scenarios where inflation occurs exclusively at 0 or 1 while retaining its core structure and flexibility. For example, in the case of only a 0-hurdle, the probability mass at one can be set to zero, and the model can focus on capturing the excess of 0s while modeling the continuous distribution for values in $(0, 1)$. Similarly, for only one hurdle, the probability mass at zero can be set to zero, and the model can concentrate on modeling the excess of 1's alongside the continuous distribution for values in $(0, 1)$. The inferential treatments for these modified cases remain mainly analogous to those discussed in this section. Estimated procedures, likelihood functions, and diagnostic tools can be applied with minimal adjustments, ensuring consistency and robustness in different scenarios. However, to maintain conciseness and avoid unnecessarily extending the work, we have chosen to omit the detailed development of these analogous cases. This decision is guided by the understanding that the methodological framework and insights provided here can be readily extended to handle these specialized cases as needed.

4.4 Data Analysis

This section uses CD34 + data, which involves 239 patients at the Edmonton Hematopoietic Stem Cell Lab at Cross Cancer Institute - Alberta Health Services from 2003 to 2008. Data can be obtained from the R software package `simpexreg` [128]. The data consists of five variables in columns. However, we only use the variable in the fourth column that corresponds to the adjusted age “ageadj” of the patients who underwent chemotherapy at the cancer institute. To obtain a set of data on $[0, 1]$,

we divided the adjusted age data by the maximum adjusted age value (i.e., 31), and henceforth, we shall refer to these scaled data as “the age data of the 239 observations in the age data; 48 of them are either 0 or 1. The frequencies of 0’s and 1’s in the age data are 46 and 2, respectively. Therefore, the data are considered to be 0 and 1 inflated; hence, it can be suitably modeled by the UIGHurdle and UUTWHurdle models. Three distinct data sets were generated using an R package to represent varying distributions of data types, specifically focusing on the proportion of zeros, ones, and fractional values (i.e., values between zero and one). These data sets were designed to evaluate the performance of the model under different data compositions. Dataset 1 was constructed with 50% zeros, 25% ones, and 25% fractional values. This configuration represents a scenario dominated by zeros, with equal contributions from ones and fractional parts. In contrast, Dataset 2 consisted of 25% zeros, 50% ones, and 25% fractional values, reflecting a case where ones are the most prevalent. Finally, data set 3 was generated with 25% zeros, 25% ones, and 50% fractional values, creating a more continuous data set dominated by fractional values. These varying compositions allow for comprehensive testing of model sensitivity to different data structures. Figures 12 to 23 present the histograms and probability-probability (P–P) plots for the three different models applied to the age data set, corresponding to Dataset 1, Dataset 2, and Dataset 3, respectively. These visualizations provide insight into how well each model captures the empirical distribution under varying proportions of zeros, ones, and fractional values. We compare the three fitted distributions of UIGHurdle1, UIGHurdle2, and UUTWHurdle distributions.

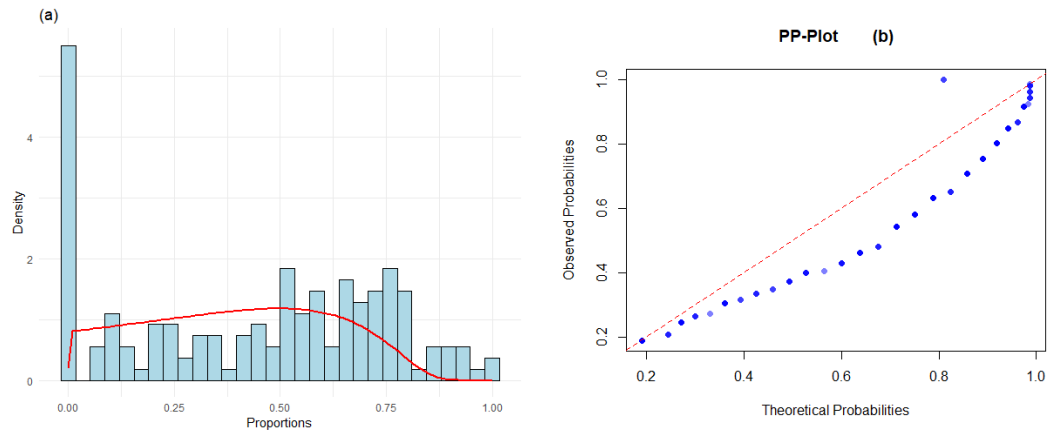


Figure 4.5: **Histogram with superimposed density curve and P–P plot of the fitted UIG-Hurdle distribution for the first transformation of the age data.**

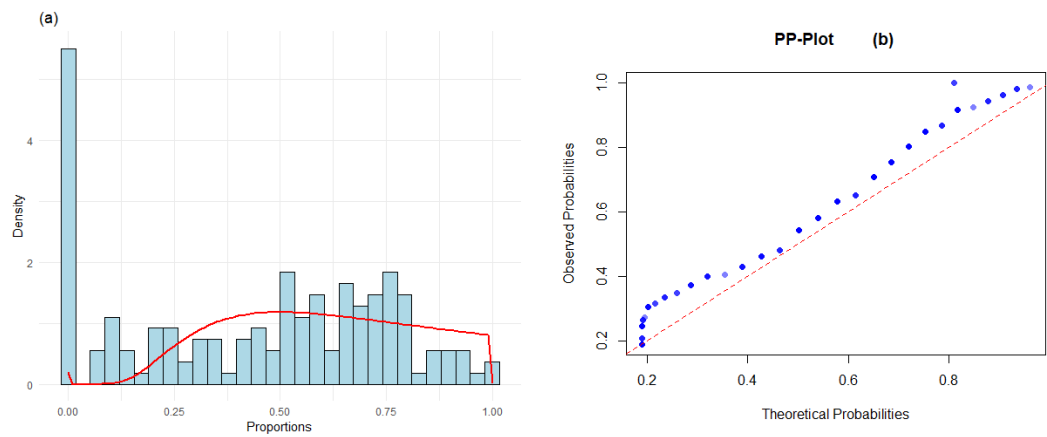


Figure 4.6: **Histogram with superimposed density curve and P–P plot of the fitted UIG-Hurdle distribution for the second transformation of the age data.**

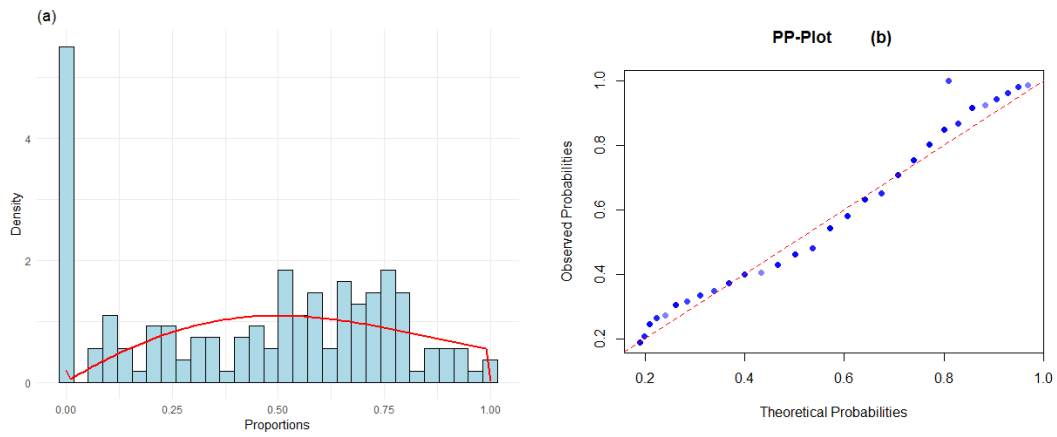


Figure 4.7: Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution of the age data.

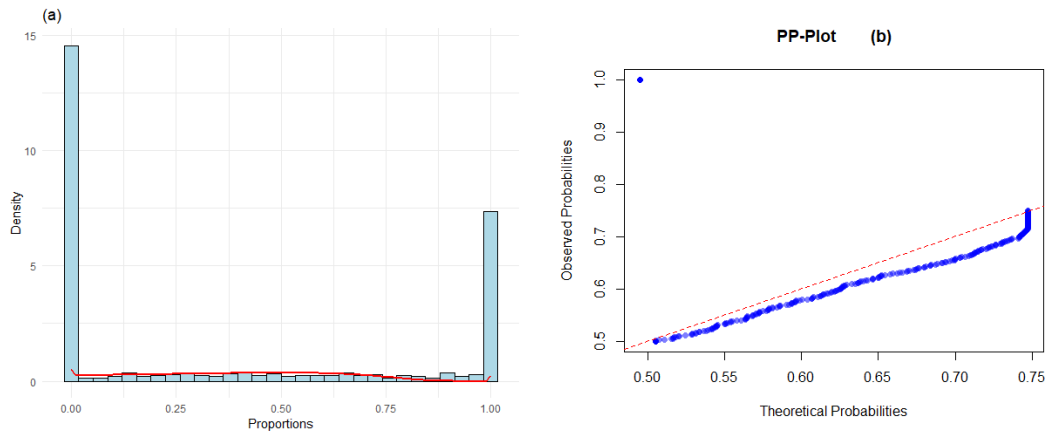


Figure 4.8: Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the first transformation of data set one.

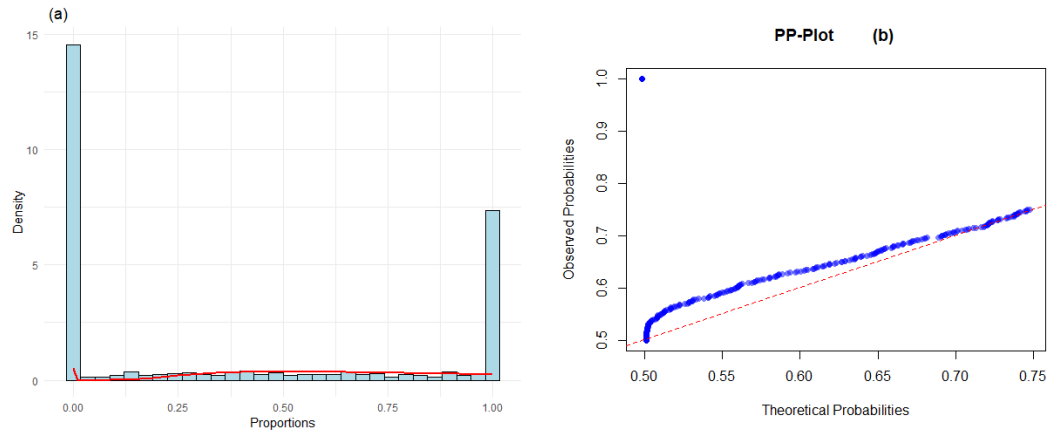


Figure 4.9: **Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the second transformation of data set one.**

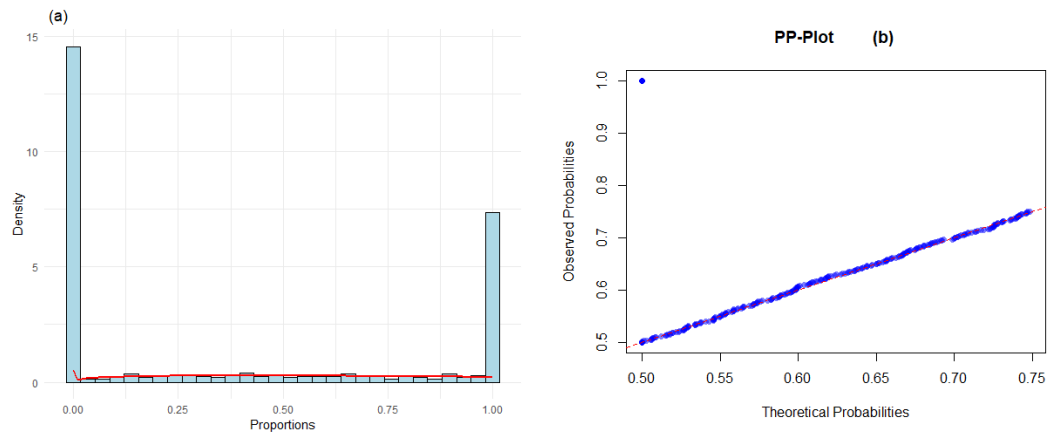


Figure 4.10: **Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution of dataset one.**

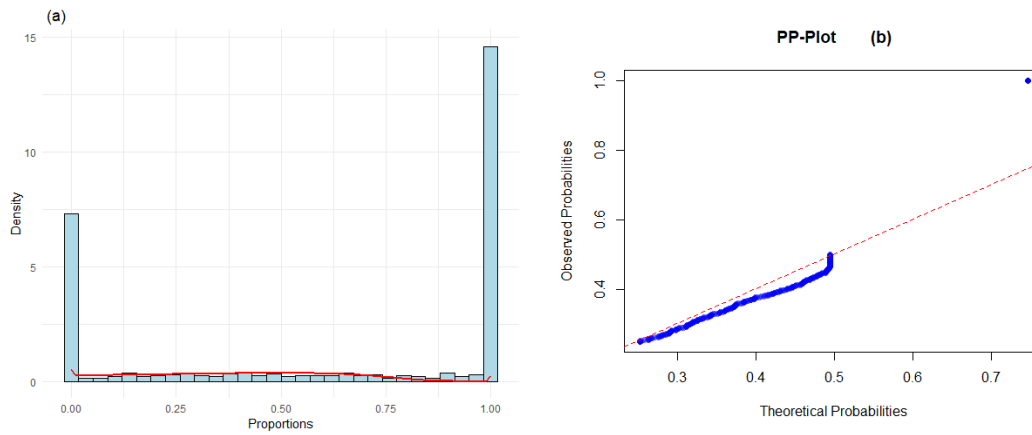


Figure 4.11: Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the first transformation of data set two.

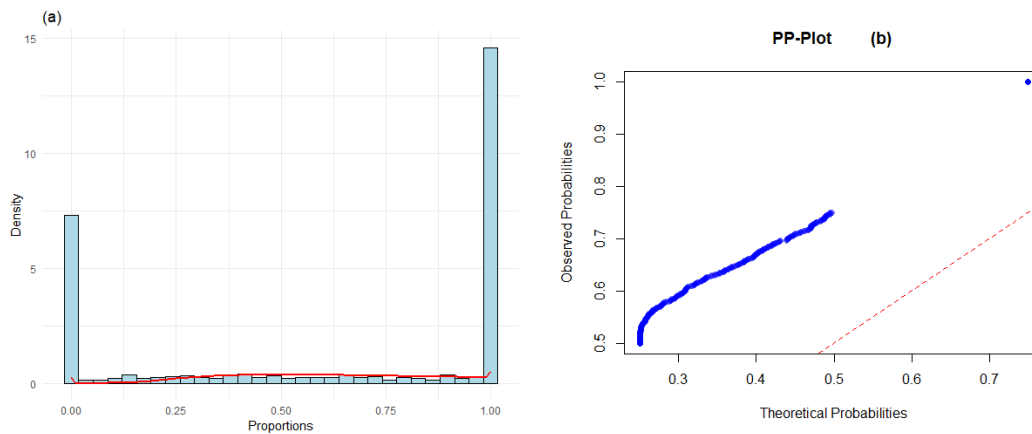


Figure 4.12: Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the second transformation of data set two.

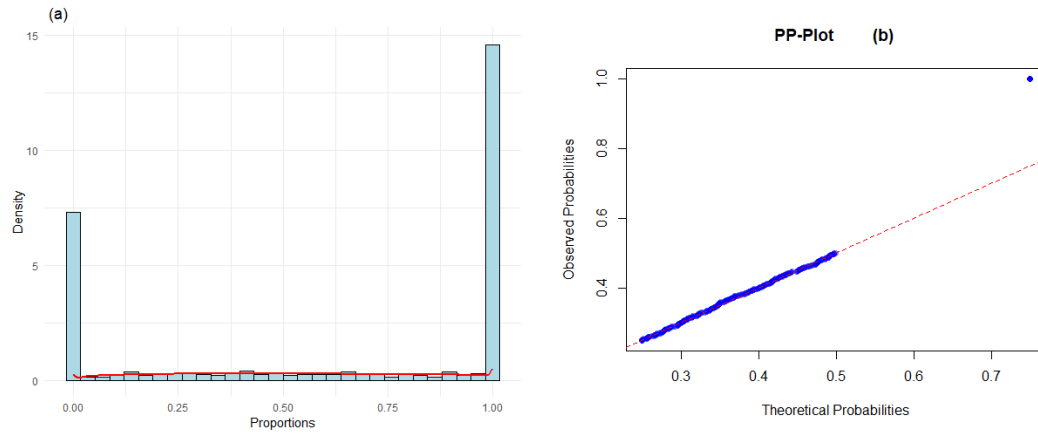


Figure 4.13: Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution of dataset two.

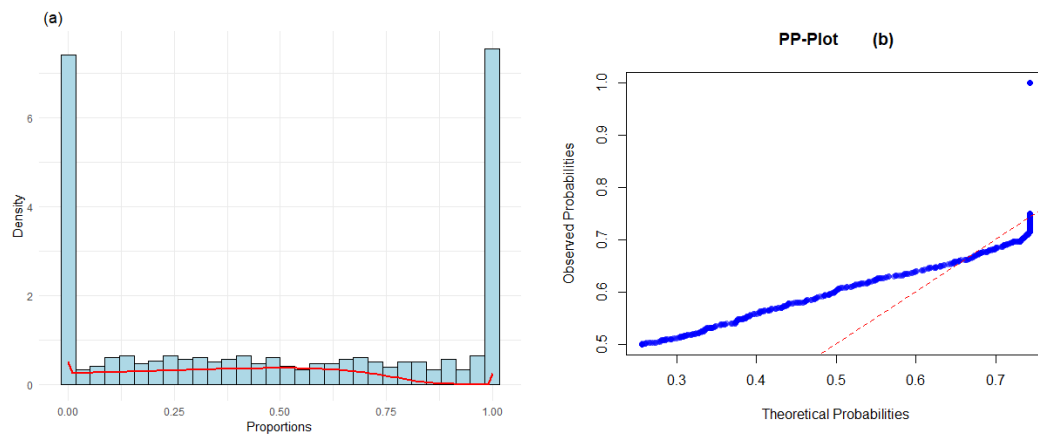


Figure 4.14: Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the first transformation of dataset three.

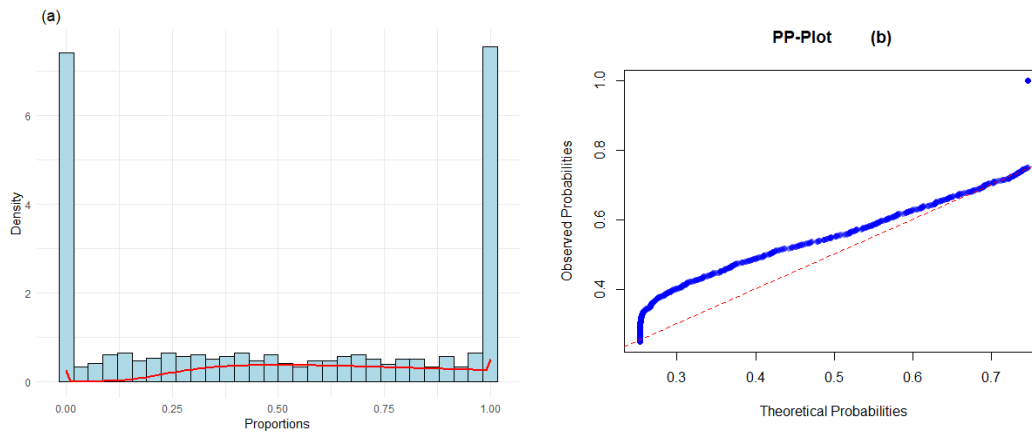


Figure 4.15: Histogram with superimposed density curve and P–P plot of the fitted UIGHurdle distribution for the second transformation of dataset three.

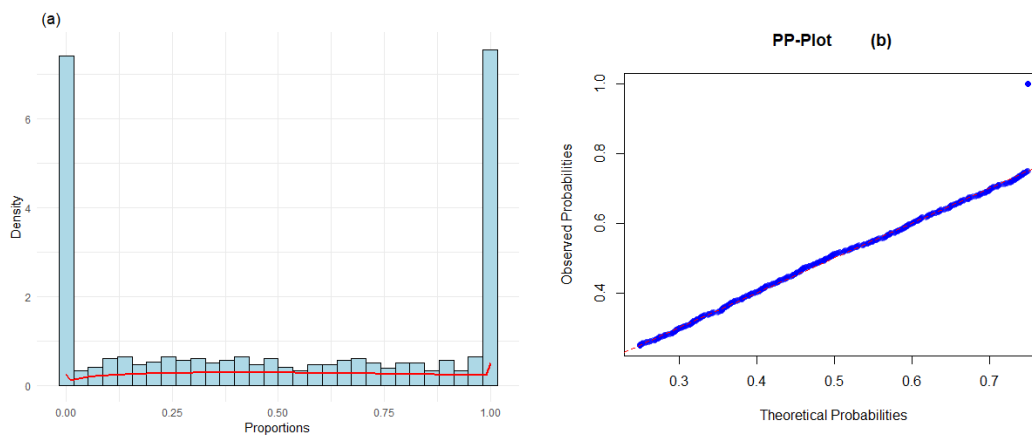


Figure 4.16: Histogram with superimposed density curve and P–P plot of the fitted UUTWHurdle distribution for the first transformation of data set three.

4.5 Estimated Result and Discussion

Table 4.1: **Fitted Results for Different 0 & 1 Hurdle Models for the Dataset**

1

Distribution(s)	Parameter Estimate				Log-likelihood & Information Criteria			
	[SE]							
	$\hat{\phi}$	$\hat{\psi}$	$\hat{\beta}$	$\hat{\lambda}$	$-\mathcal{L}$	AIC	BIC	AICc
UIGHurdle1	0.7583409	0.3333337	0.0100000	-	-1377.393	2760.786	2775.509	2760.810
	$[1.35 \times 10^{-2}]$	$[1.72 \times 10^{-2}]$	[-]					
UIGHurdle2	0.7530112	0.3333338	0.0100000	-	-1235.218	2476.435	2491.159	2479.46
	$[1.36 \times 10^{-2}]$	$[1.72 \times 10^{-2}]$	[-]					
UUTWHurdle	0.7499995	0.3333316	1.3149713	0.7756274	-1037.002	2082.004	2101.635	2082.044
	$[1.37 \times 10^{-2}]$	$[1.72 \times 10^{-2}]$	$[0.13185]$	$[0.39052]$				

Table 4.2: **Fitted Results for Different 0 & 1 Hurdle Models for Dataset 2**

Distribution(s)	Parameter Estimate				Log-likelihood & Information Criteria			
	[SE]							
	$\hat{\phi}$	$\hat{\psi}$	$\hat{\beta}$	$\hat{\lambda}$	$-\mathcal{L}$	AIC	BIC	AICc
UIGHurdle1	0.7583398	0.6666663	0.0100000	-	-1377.394	2760.788	2775.511	2760.812
	$[1.36 \times 10^{-2}]$	$[1.72 \times 10^{-2}]$	[-]					
UIGHurdle2	0.7530109	0.6666657	0.0100000	-	-1235.005	2476.437	2491.161	2476.462
	$[1.36 \times 10^{-2}]$	$[1.72 \times 10^{-2}]$	[-]					
UUTWHurdle	0.7499993	0.6666680	1.3149255	0.7755160	-1037.005	2082.01	2101.641	2082.050
	$[1.37 \times 10^{-2}]$	$[1.72 \times 10^{-2}]$	$[0.13185]$	$[0.39053]$				

Table 4.3: Fitted Results for Different 0 & 1 Hurdle Models for Dataset 3

Distribution(s)	Parameter Estimate				Log-likelihood & Information Criteria			
	$\hat{\phi}$	$\hat{\psi}$	$\hat{\beta}$	$\hat{\lambda}$	$-\mathcal{L}$	AIC	BIC	AICc
UIGHurdle1	0.5138746	0.500000	0.0100000	-	-1765.394	3536.787	3551.511	3536.812
	[1.62×10 ⁻²]	[2.24×10 ⁻²]	[5.27×10 ⁻³]					
UIGHurdle2	0.7530109	0.500002	0.0100000	-	-1580.636	3167.272	3181.995	3167.296
	[1.59×10 ⁻²]	[2.23×10 ⁻²]	[5.06×10 ⁻³]					
UUTWHurdle	0.5000000	0.5000000	1.1291979	0.4308606	-1038.607	2085.214	2104.846	2085.255
	[1.58×10 ⁻²]	[2.24×10 ⁻²]	[8.76×10 ⁻²]	[0.28568]				

Table 4.4: Fitted Results Hurdle Models for the 0 and 1 Inflated Age Data.

Distribution(s)	Parameter Estimate				Log-likelihood & Information Criteria			
	$\hat{\phi}$	$\hat{\psi}$	$\hat{\beta}$	$\hat{\lambda}$	$-\mathcal{L}$	AIC	BIC	AICc
UIGHurdle1	0.20382341	0.06250449	0.0100000	-	-145.3549	296.7097	305.8975	296.8656
	[3.21×10 ⁻²]	[4.28×10 ⁻²]	[1.41×10 ⁻²]					
UIGHurdle2	0.20253285	0.06250517	0.0100000	-	-134.223	274.446	283.6338	274.6018
	[3.20×10 ⁻²]	[4.28×10 ⁻²]	[1.32×10 ⁻²]					
UUTWHurdle	0.20253270	0.06249981	1.92061686	1.82659177	-79.21484	166.4297	178.6801	166.6911
	[3.20×10 ⁻²]	[4.28×10 ⁻²]	[0.2343]	[0.5757]				

The tables present the parameter estimates and model fit statistics for different 0 and 1 hurdle distributions applied to four datasets: three synthetic datasets (Tables 25–27) and one real-world age dataset (Table 28). The models evaluated include UIGHurdle1, UIGHurdle2, and UUTWHurdle. In all datasets, the UUTWHurdle model consistently demonstrates a better fit, as indicated by lower log-likelihood $-\mathcal{L}$, AIC, BIC, and AICc values. For example, in Dataset 1 (Table 25), UUTWHurdle

yields the lowest AIC (2082.004) compared to UIGHurdle1 (2760.587) and UIGHurdle2 (2476.435). Similar trends are observed in datasets 2 and 3 (Tables 26 and 27), where the UUTWHurdle model outperforms its counterparts in terms of model fit criteria, indicating its robustness in handling varying proportions of zeros, ones, and fractional values.

The performance of the models is further validated using the real-world age dataset (Table 28), where the UUTWHurdle model again exhibits the most favorable fit with a significantly lower AIC (166.4297), BIC (178.6801), and AICc (166.6911), as well as the least negative log-likelihood (-79.2148). In contrast, the UIGHurdle1 and UIGHurdle2 models produce notably higher AIC and BIC values, suggesting a poorer fit to the empirical data. These findings underscore the superior flexibility and accuracy of the UUTWHurdle model in capturing the data structure, particularly in the presence of both zero and one inflation. This makes it a more suitable candidate for modeling datasets characterized by boundary inflation and complex underlying distributions.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Chapter 2 provides a comprehensive overview of Generalized Linear Models (GLMs), detailing their components, including the random component, systematic component, and link function. It explains how GLMs extend traditional linear models to accommodate various response variables, such as counts, binary outcomes, and proportions. The chapter also discusses parameter estimation techniques, such as Iteratively Reweighted Least Squares (IRLS), and hypothesis testing methods like the analysis of deviance. The Poisson regression model is highlighted as a key tool for modeling count data, with an illustrative example demonstrating its application.

Chapter 3 provides a comprehensive analysis of smoking prevalence among the entire population across different locations and regions, highlighting the variations in smoking likelihood among various environments. The findings indicate that smok-

ing behavior is significantly influenced by contextual and geographical factors, with coffee shops and public places emerging as the most prominent locations for smoking. Notably, smoking in coffee shops is more likely compared to smoking in cars, while public places exhibit a higher smoking prevalence than in cars. The absence of significant differences between smoking at home and in cars suggests that domestic environments may not play a central role in determining smoking behaviors. SHIS 2014 [1] survey reported that 17.2 % of Saudis were exposed to smoking at home, while 14.1% were exposed in the workplace. A detailed analysis of smoking prevalence among the male population across different locations and regions offers insights into variations in smoking likelihood in specific environments.

The findings indicate that smoking behaviors differ significantly based on the place of exposure, with coffee shops and public places emerging as the most prominent locations for smoking. Specifically, smoking in coffee shops is more likely than in cars, while public places exhibit the highest smoking prevalence, being more likely than in cars. Conversely, smoking at home is less prevalent than in cars, indicating a relatively reduced likelihood of smoking in domestic environments. A comprehensive analysis of smoking prevalence among the female population across different regions and locations highlights significant disparities in smoking behaviors. The findings indicate that public places and coffee shops are the most common locations for smoking among women, with public places showing a higher likelihood of smoking than cars.

Similarly, coffee shops exhibit a significantly higher smoking prevalence, being more likely than cars. In contrast, smoking at home is slightly less prevalent, with odds 1.13 times lower than in cars, suggesting that indoor smoking may be somewhat restricted

among females. The likelihood of smoking in workplaces is relatively moderate, higher than in cars, indicating that occupational environments play a role in shaping smoking behaviors. [6] reported that males smoke more than females, indicating a higher prevalence of smoking among men. Furthermore, Madinah is less likely to have smoking occurrences. Other regions, including Al-Baha, Al-Jouf, the Eastern Region, Hail, Jazan, Najran, the Northern Borders, Qaseem, Tabouk, and the Western Region, also show lower smoking prevalence. However, our results are partially consistent with [6] as Makkah, Riyadh, and Hail exhibit higher smoking rates.

These findings suggest that regional socio-cultural norms and regulatory frameworks may contribute to variations in smoking behaviors. Furthermore, location-based analysis within regions reveals distinct patterns. For instance, in Al-Baha, smoking in coffee shops is considerably less likely than smoking in a car, whereas smoking at home is notably more prevalent. A similar trend is observed in the Eastern region, where tobacco at home, in public places, and in workplaces is significantly higher compared to smoking in cars. This pattern is evident in several other areas, including Hail, Jazan, and Al-Jouf, demonstrating the context-dependent nature of smoking behaviors.

In the male population, smoking in Al-Baha, the Eastern Region, Hail, Al-Jouf, Al-Madinah, Najran, the Northern Borders, and Al-Qassim is significantly less likely compared to Aseer, whereas Makkah, Riyadh, and Tabouk exhibit higher smoking prevalence, with Makkah showing the highest likelihood, more than in Aseer. The lack of significant differences in Jazan suggests that smoking patterns in that region may be more evenly distributed across various locations. These results underscore the socio-

cultural and regulatory differences across regions, which may impact smoking behavior and exposure patterns. A deeper location-based analysis within regions reveals notable variations. For instance, in Al-Baha, smoking in coffee shops is significantly less likely than in cars, while smoking at home is notably more prevalent. A similar trend is observed in the Eastern Region, where tobacco consumption at home, in public places, and in workplaces is considerably higher than that of cars. These patterns suggest that the availability of designated smoking areas, regulatory enforcement, and cultural norms contribute to regional differences in smoking behavior.

The findings highlight substantial regional disparities in smoking prevalence among the female population, with some areas exhibiting a higher likelihood of smoking. In contrast, others show a significant decline compared to Aseer. Smoking prevalence is significantly lower in Al-Baha, the Eastern Region, Hail, Jazan, Al-Jouf, Al-Madinah, Najran, the Northern Borders, and Al-Qassim than in Aseer. However, Makkah, Riyadh, and Tabouk exhibit a notably higher smoking prevalence. Makkah shows the highest odds at 3.32 times more than Aseer, followed by Riyadh at 2.59 times and Tabouk at 2.29 times. These variations suggest that socio-cultural norms, economic activity, urbanization, and regulatory policies play a critical role in influencing smoking habits among females. Further location-based analysis within regions reveals unique smoking behaviors among women in different geographical areas.

For example, in Al-Baha, smoking in coffee shops and public places is significantly less likely than in cars, whereas smoking at home is notably more prevalent. A similar pattern is observed in the Eastern Region, where home smoking is substantially higher than in cars, while public and workplace smoking remains significantly elevated. The

urban-rural divide is evident in these findings, with urbanized regions such as Makkah and Riyadh exhibiting higher public smoking prevalence. In contrast, rural areas experience higher smoking rates in domestic settings.

Comparison of males and females: This study provides a comprehensive comparative analysis of smoking likelihood among males and females across different regions and smoking locations, offering critical insights into gender-specific smoking behaviors. The findings highlight that public places and coffee shops remain the most prevalent smoking locations for both genders, although subtle gender-based variations exist in regional smoking patterns. Public places exhibit the highest smoking prevalence among both males and females, compared to smoking in cars. Coffee shops remain a common smoking location, with both males and females having nearly identical odds of smoking in these settings. Smoking at home presents notable gender differences, with females showing a slightly higher likelihood compared to males showing less likelihood.

Workplaces demonstrate a higher smoking likelihood for males than females, indicating that occupational environments play a more significant role in male smoking behavior. Regions with lower smoking prevalence compared to Aseer are Al-Baha, the Eastern Region, Hail, Jazan, Al-Jouf, Al-Madinah, Najran, the Northern Borders, and Al-Qassim, which show reduced smoking likelihoods for both genders. Regions with higher smoking prevalence compared to Aseer are Makkah, Riyadh, and Tabouk, which exhibit greater smoking prevalence for both genders. Males exhibit higher odds of smoking in Riyadh and Makkah compared to Aseer, whereas females show slightly lower but still elevated odds in Riyadh and Makkah. Tabouk exhibits a unique pattern

where females are more likely to smoke than males.

In Al-Baha, smoking at home is more prevalent for both genders, with females slightly more likely than males. Conversely, smoking in public places is significantly less likely for both genders, but females exhibit a more substantial reduction than males. In the Eastern Region, smoking at home is substantially more likely for both genders, with females significantly outpacing males. However, smoking in coffee shops is more likely for males but less likely for females. Hail gender differences are pronounced: males are less likely to smoke in coffee shops, while females are significantly more likely. At home, females are more likely to smoke than males. In Jazan, smoking at home is more likely for both genders, but females exhibit much stronger odds compared to males. Smoking in workplaces is significantly less likely for females compared to males.

In Al-Jouf, smoking in coffee shops is less likely for both genders, with females showing a more significant reduction compared to males. However, smoking at home is more likely for both genders, with females showing higher odds than males. In Makkah, smoking at home is more likely for both genders, with females showing higher odds than males. However, smoking in public places is more likely for males but less likely for females.

In Najran, males are more likely to smoke in coffee shops, while females are less likely. Smoking at home is more likely for both genders, with females slightly exceeding males. In Northern Borders, smoking is less likely in coffee shops for both genders, but females exhibit a much more substantial reduction than males. Smoking in public places follows a similar pattern, where males are more likely to smoke while females

are less likely. In Al-Qassem, smoking in coffee shops is less likely for both genders, with females showing a more significant reduction than males. However, smoking at home is more likely for both genders, with females exhibiting more substantial odds compared to males.

In Riyadh, smoking is less likely in coffee shops for both genders, but females show a more significant reduction compared to males. Smoking at home is more likely for both genders, with females exhibiting slightly stronger odds than males. In Tabouk, smoking in coffee shops is much less likely for both genders, with females exhibiting a much more substantial reduction than males. However, smoking at home is more likely for both genders, with females surpassing males.

Comparison of males and the entire population: This study provides a comparative analysis of smoking likelihood between males and the entire population across various regions and smoking environments. Public places and coffee shops remain the most frequent smoking locations for both males and the whole population. The likelihood of smoking in public places is nearly identical for males and the entire population compared to smoking in cars. Similarly, smoking in coffee shops is more likely for males and the whole population, indicating a consistent trend across both groups. Workplaces show a higher smoking likelihood for males than for the entire population, suggesting that occupational environments play a more significant role in male smoking behavior.

Smoking at home presents notable differences, and the whole population shows no significant difference between smoking at home and in a car. However, males are less likely to smoke at home than in a car, suggesting gender-specific behavioral distinctions

in home smoking habits.

Regions where smoking is less likely compared to Aseer are Al-Baha, the Eastern Region, Hail, Jazan, Al-Jouf, Al-Madinah, Najran, the Northern Borders, and Al-Qassem, which show a lower likelihood of smoking for both males and the entire population. For instance, in Al-Baha, smoking is less likely for the whole of the population and less likely for males compared to Aseer. Regions where smoking is more likely compared to Aseer are Makkah, Riyadh, and Tabouk, which exhibit a higher smoking prevalence, with males having slightly higher odds than the entire population in Riyadh and Makkah. In Makkah, smoking is more likely to be done by males and more likely for the whole of the population, highlighting a strong regional effect. Males show no significant difference in smoking prevalence between Jazan and Aseer, whereas the entire population is less likely to smoke in Jazan compared to Aseer. In Al-Jouf, males exhibit a lower smoking likelihood compared to the entire population, suggesting a more pronounced reduction for males. In Al-Baha, smoking in coffee shops is less likely for both males and the entire population. Smoking at home is more likely for both males and the entire population. In the Eastern Region, males are more likely to smoke in coffee shops, while the entire population is less likely. Smoking at home is significantly higher for both males and the entire population. In Hail, smoking in coffee shops is less likely for males and the entire population. Smoking at home is significantly more likely for both males and the entire population. In Makkah, smoking in coffee shops is less likely for males and the entire population. Smoking in public places is more likely for males but less likely for the entire population. In Najran, males are more likely to smoke in coffee shops, while the entire population

shows a slightly lower likelihood. Smoking at home is more likely for both males and the entire population.

Comparison of females and the entire population: This study provides a comparative analysis of smoking likelihood between females and the entire population across various regions and smoking environments. Public places and coffee shops remain the most frequent smoking locations for both females and the whole population. The likelihood of smoking is more likely in public places, and it's nearly identical for both females and the entire population compared to smoking in cars. Similarly, smoking in coffee shops is more likely for both females and the whole population, indicating a consistent trend. Workplaces show a higher smoking likelihood for females than for the entire population, suggesting that workplace smoking is a more significant concern for females relative to the overall population. Smoking at home presents notable gender differences. The whole population shows no difference between smoking at home and in a car. However, females are more likely to smoke at home than in a car, suggesting a higher prevalence of indoor smoking among women.

Regions where smoking is less likely compared to Aseer are Al-Baha, the Eastern Region, Hail, Jazan, Al-Jouf, Al-Madinah, Najran, the Northern Borders, and Al-Qassem, which show a lower likelihood of smoking for both females and the entire population. For instance, in Al-Baha, smoking is less likely for both the entire population and females compared to Aseer. Regions where smoking is more likely compared to Aseer are Makkah, Riyadh, and Tabouk, which exhibit a higher smoking prevalence, with the entire population having slightly higher odds than females in Riyadh and Makkah. In Makkah, smoking is more likely to be found among females

and the entire population, indicating a strong regional effect. Both females and the entire population are less likely to smoke in Jazan than in Aseer. In Al-Jouf, females exhibit a lower smoking likelihood compared to the entire population, suggesting a more pronounced reduction for females. In Al-Baha, smoking in coffee shops is less likely for females and the entire population. Smoking at home is more likely for both females and the entire population. In the Eastern Region, females are less likely to smoke in coffee shops, whereas the entire population shows a slightly higher likelihood. Smoking at home is significantly higher for both females and the entire population. In Hail, females are more likely to smoke in coffee shops, while the entire population is less likely. Smoking at home is significantly more likely for both females and the entire population. In Makkah, smoking in coffee shops is less likely for females and the entire population. Smoking in public places is less likely for females but more likely for the entire population. In Najran, females are less likely to smoke in coffee shops, while the entire population is more likely. Smoking at home is more likely for both females and the entire population.

This study provides an in-depth analysis of smoking likelihood across different age groups, highlighting significant variations in smoking behavior and location preferences within the entire population. The possibility of smoking decreases progressively with age, with individuals aged 65+ being less likely to smoke compared to the 10-14 age group. Similarly, individuals aged 60-64, 55-59, and 50-54 also show a decline in smoking prevalence compared to the youngest age group. Individuals aged 20-24 exhibit the highest smoking likelihood compared to 10-14-year-olds, followed closely by the 15-19 and 25-29 age groups. These results suggest that early adulthood is a

critical period for tobacco initiation and sustained smoking behaviors, necessitating stronger prevention measures targeting youth and young adults. Smoking in public places is significantly more prevalent. The likelihood of smoking in public places and at home is higher compared to smoking in a car, making it the most common smoking environment. In coffee shops, smoking is less likely compared to smoking in a car. Workplace smoking is comparatively less likely than in cars, possibly due to workplace smoking restrictions and policies. The odds of smoking in coffee shops among individuals aged 20-24 are higher than in cars, highlighting a strong preference for social smoking environments. For individuals aged 25-29, the likelihood remains high compared to cars, reinforcing the role of coffee shops as a familiar smoking venue for young adults. Middle-aged and older individuals (40+) exhibit a declining preference for coffee shops. Individuals aged 40-44 are more likely to smoke in coffee shops than in cars, but this likelihood declines. For individuals aged 50-54, the possibility of smoking in coffee shops is higher, which further reduces to 1.77 times for individuals aged 65+. Public places remain the most common smoking locations for all age groups. Both Individuals aged 45-49 and 50-54 are more likely to smoke in public places. The likelihood remains consistently high across age groups, reinforcing the need for more vigorous enforcement of public smoking bans and designated smoking zones. Older individuals (60+) show a reduced likelihood of smoking across all locations. Individuals aged 60-64 are more likely to smoke in coffee shops compared to younger groups with higher smoking rates. The odds of smoking in coffee shops for individuals aged 65+ are significantly lower than those of middle-aged individuals, suggesting a gradual decline in social smoking with age.

This study provides an in-depth analysis of smoking prevalence among males across various age groups, examining age-specific variations in smoking behavior and preferred locations. Smoking prevalence declines significantly with increasing age. Older males (65+) are less likely to smoke than those in the 10-14 age group, demonstrating a substantial decline in smoking prevalence with age. Similarly, males aged 60-64, 55-59, and 50-54 show a progressive reduction in smoking likelihood, suggesting that smoking cessation efforts are more effective among older populations. Peak smoking prevalence is observed among young adults (15-24 years old). Males aged 20-24 are more likely to smoke, followed by 15-19-year-olds and 25-29-year-olds, compared to the 10-14 age group. These findings underscore the need for stronger tobacco control measures aimed at adolescents and young adults, including school-based prevention programs and media campaigns targeting early smoking initiation. Public places have emerged as the most prevalent smoking locations. Smoking in public places is more likely compared to smoking in cars, making it the most common location for male smokers. Home smoking prevalence is also high, highlighting the risk of secondhand smoke exposure for family members. Coffee shops remain a popular smoking venue, particularly for young and middle-aged adults. Workplace smoking is relatively lower, possibly due to strict workplace regulations and smoking bans.

The likelihood of smoking in coffee shops is highest among males aged 20-24 compared to cars, reinforcing the role of social interactions in smoking habits. Similarly, public places are preferred locations for young smokers, with males aged 20-24 exhibiting a higher likelihood of smoking in public spaces than in cars. Middle-aged and older males (40+) show a declining probability of smoking in coffee shops. Males aged

40-44 are more likely to smoke in coffee shops, but this preference declines for older age groups. By contrast, males aged 60-64 are more likely to smoke in coffee shops, indicating a shift away from social smoking environments as age increases. Workplace Smoking Increases with Age, Reaching Its Peak at 40-54 Years Old. The likelihood of workplace smoking is significantly higher among middle-aged males. Both Males aged 35-39 and 40-44 are more likely to smoke in the workplace. This trend suggests that workplace smoking remains a concern for middle-aged working men, emphasizing the need for stronger workplace smoking policies and designated smoking zones. Smoking likelihood in the workplace declines after age 55. For males aged 55-59, the likelihood of smoking in the workplace is higher and further decreases for males aged 60-64. This indicates that smoking cessation efforts may be more effective in older working populations, possibly due to health concerns and medical interventions.

This study presents an in-depth examination of smoking prevalence among females across different age groups, providing insights into age-related smoking behaviors and preferences for smoking locations. Females aged 65+ are less likely to smoke compared to those in the 10-14 age group, highlighting a significant decline in smoking likelihood with age. A similar downward trend is observed among the 60-64, 55-59, and 50-54 age groups, reinforcing the pattern of reduced smoking behavior among older populations. Smoking prevalence is highest among middle-aged females (25-39 years old). The highest smoking likelihood is observed in the 25-29 age group compared to the 10-14 age group, indicating that smoking is more prevalent among younger and middle-aged women. Females aged 35-39 and 30-34 show higher smoking prevalence, suggesting that smoking behaviors peak in early adulthood. This trend underscores

the importance of targeted smoking prevention programs for young and middle-aged females. Smoking in public places is more likely than smoking in a car, making it the most frequent location for female smokers. Home smoking is also prevalent, indicating a higher risk of secondhand smoke exposure for family members. Coffee shops are more likely to serve as a key social smoking environment, particularly among younger women. Workplace smoking is the least common, potentially due to more stringent workplace smoking policies and cultural norms. Coffee shop smoking, home smoking, public place smoking, and workplace smoking all exhibit slight increases for females aged 15-19 compared to 10-14-year-olds.

These findings suggest that smoking initiation occurs during adolescence, reinforcing the need for school-based smoking prevention programs. Smoking likelihood in all locations is nearly unchanged, with coffee shop smoking, home smoking, and public place smoking showing only slight increases compared to 10-14-year-olds. This suggests that smoking rates stabilize among young adult women, potentially due to social and environmental factors. Smoking in coffee shops declines significantly for the 25-29 age group, suggesting that social smoking declines with age. Conversely, home smoking prevalence remains high, with 25-29-year-olds, 30-34-year-olds, and 35-39-year-olds showing continued home smoking tendencies. Workplace smoking remains relatively low among this demographic, emphasizing the role of workplace smoking restrictions. Females aged 40-44 show an increase in coffee shop smoking and public place smoking, suggesting a resurgence in social smoking behaviors. However, home and workplace smoking prevalence remains stable, indicating that smoking cessation efforts may be more effective among older females. Workplace smoking is highest

among females aged 50-54; females aged 50-54 are more likely to smoke in the workplace, making this the peak smoking age for working women. This highlights the need for stronger workplace smoking cessation programs and designated smoking areas. Workplace smoking remains relatively low for younger females. Among females aged 15-24, workplace smoking likelihood is only slightly higher, possibly due to workplace smoking bans and social restrictions. Smoking prevalence in the workplace declines after age 55. Both females aged 55-59 and 60-64 are more likely to smoke in the workplace. By age 65+, workplace smoking prevalence is higher, indicating a natural decline in smoking among older working women.

Comparison of age males and females, this study provides a comprehensive analysis of smoking prevalence among males and females across different age groups, emphasizing gender disparities in smoking behaviors and location preferences. Smoking likelihood declines significantly in older age groups for both genders. Both males and females aged 65+ are considerably less likely to smoke compared to the 10-14 age group, confirming a general decline in smoking prevalence with age. A similar decrease is observed in the 60-64, 55-59, and 50-54 age groups, reinforcing the trend of reduced smoking among older individuals. Smoking is most prevalent among young and middle-aged males. Males in the 20-24 age group exhibit the highest smoking likelihood compared to the 10-14 age group, whereas females in the same age group are less likely to smoke. Similarly, males aged 15-19 are more likely to smoke, while females in the same age group are less likely to smoke. This indicates that smoking initiation occurs earlier among males, while females show a delayed or reduced uptake of smoking behavior. Smoking prevalence remains higher among middle-aged males

compared to females. Males aged 35-39 and 30-34 show higher smoking prevalence compared to the 10-14 age group.

In contrast, females aged 35-39 and 30-34 show lower smoking prevalence in comparison. This suggests that middle-aged males are more likely to sustain smoking habits, whereas females may exhibit more smoking cessation behaviors or social limitations influencing their smoking habits. Public places remain the most common smoking location for both genders. Males and females predominantly smoke in public places, confirming public smoking as a key area for smoking interventions. Home smoking is more prevalent for both genders. Males and females are significantly more likely to smoke at home compared to in a car, indicating higher secondhand smoke exposure risks within households. Coffee shop smoking prevalence is similar across genders. Both males and females show a slightly lower likelihood of smoking in coffee shops compared to cars, suggesting moderate social smoking behaviors. Workplace smoking is more common among males. Males are less likely to smoke in the workplace compared to cars, while females exhibit an even lower workplace smoking prevalence. This indicates that males have higher workplace smoking rates than females, possibly due to occupational environments that accommodate smoking behaviors. Teenage males (15-19) are likelier to smoke in coffee shops, while females prefer public places. Both males and females aged 15-19 are more likely to smoke in coffee shops. However, public place smoking is lower for males but higher for females, suggesting that females are more likely to smoke in open social environments.

Males in this age group (15-19) are significantly more likely to smoke in coffee shops compared to females. Workplace smoking is also considerably higher among males

than females, reflecting gendered differences in occupational settings and smoking habits. Males aged 35-39 are more likely to smoke in workplaces, whereas females in the same age group are less likely to do so. Similarly, males aged 30-34 are more likely to smoke in workplaces, while females in this age group are less likely to do so. These findings suggest that workplace smoking interventions should prioritize male populations, particularly those in middle-aged groups. For the 50-54 age group, both males and females are more likely to smoke in workplaces. Similarly, both males and females in the 55-59 age group are more likely to smoke in workplaces. These patterns indicate that older males continue to smoke in professional settings, whereas females are more likely to reduce their smoking habits with age.

Comparison of age males and the entire population This analysis presents a detailed comparison of smoking prevalence among males versus the entire population (males and females combined) across different age groups and smoking locations. Smoking decreases with age in both groups, but males have consistently higher smoking rates. For ages 65+, 60-64, 55-59, and 50-54, the odds of smoking are lower for both males and the entire population compared to the 10-14 age group. For ages, the odds are 0.31 times lower for males and 0.26 times for the whole population. The entire population exhibits slightly higher odds of smoking in older age groups than males alone, indicating that females in these age brackets contribute to the overall smoking rate despite being a smaller proportion of smokers. Smoking prevalence peaks in young adults (20-24) for both groups but is significantly higher among males. Both males and the entire population are more likely to smoke in the 20-24 age group relative to the 10-14 age group. For the 15-19 age group, both males and the en-

tire population are more likely to smoke, indicating a greater likelihood of smoking initiation among young males. Middle-aged adults (30-39) show sustained smoking behavior, with males leading. Both males and the entire population are more likely in the 30-34 and 35-39 age groups to show sustained smoking prevalence. This suggests that while both groups sustain smoking in middle age, males have a slightly stronger tendency to continue smoking over time.

Public places are the most common smoking location for both groups. Males and the entire population are more likely to smoke in public places compared to smoking in a car. Home smoking is similarly prevalent in both groups. Both males and the entire population are more likely to smoke at home than in a car. Males and the entire population show a slightly lower likelihood of smoking in coffee shops than in cars. Similarly, males and the entire population are less likely to smoke in workplaces than in cars. Both males and the entire population are significantly more likely to smoke in coffee shops in the 15-19, 20-24, 30-39, and 35-39 age groups. Both males and the entire population are significantly less likely to smoke at home in the 15-19, 20-24, 30-39, and 35-39 age groups. Smoking in public places is nearly the same for both groups, with a slightly lower prevalence in males and the entire population in the 15-19 and 20-24 age groups. Public place smoking remains prevalent and higher in both males and the entire population in the 30-34 and 35-39 age groups. The likelihood of smoking in workplaces for both males and the entire population is less likely in the 15-19 age group. Workplace smoking is significantly more likely for both males and the entire population in the 20-24, 35-34, 50-54, 55-59, 60-64, and 65+ age groups.

Comparison of age females and the entire population, this analysis explores smok-

ing prevalence among females compared to the entire population (males and females combined) across various age groups and smoking locations. For ages 65+, 60-64, 55-59, and 50-54, the odds of smoking are lower for females and the entire population compared to the 10-14 age group. For ages 20-24 and 15-19, females are less likely to smoke, whereas the entire population is more likely to smoke than the 10-14 age group. This suggests that smoking initiation is more common among young males, which increases overall smoking rates in the entire population. For the 35-39, 30-34, and 25-29 age groups, both females and the entire population are more likely to smoke. These results suggest that while female smoking rates are rising in middle age, they remain slightly lower than those of the entire population, likely due to the higher prevalence of male smokers in these age groups. Public places are the most common smoking location for both groups, but slightly more so in the entire population. Both females and the entire population are more likely to smoke in public places and at home compared to smoking in a car. Both females and the entire population show a slightly lower likelihood of smoking in coffee shops and workplaces compared to cars. Both females and the entire population are significantly more likely to smoke in coffee shops in the 15-19 and 20-24 age groups. Females show lower coffee shop smoking rates, whereas the entire population is more likely to smoke in the 30-34 and 35-39 age groups.

The likelihood of smoking at home for females is higher and lower for the entire population in the 15-19, 20-24, and 30-34 age groups. For the age of 35-39, the likelihood of smoking at home for both females and the entire population is less likely. The likelihood of smoking in public places for females is higher and lower for the entire

population in the 15-19, 20-24, and 30-34 age groups. For the age of 35-39, females are less likely to smoke in public places, whereas the entire population is more likely to smoke in public places. The likelihood of smoking in the workplace for females is higher and lower for the entire population in the 15-19 and 30-34 age groups. The likelihood of smoking in the workplace for both females and the entire population is more likely in the 20-24, 50-54, 55-59, 60-64, and 65+ age groups. For the age of 35-39, females are less likely to smoke in public places, whereas the entire population is more likely to smoke in public places.

In this study, the likelihood of smoking varies across the 13 administrative regions of Saudi Arabia and is consistent with the findings from a recent survey [14]. Compared to Makkah, Riyadh and Tabouk are more likely to smoke. At the same time, those in regions like Al-Baha, Hail, the Eastern region, Jazan, Al-Jouf, Najran, the Northern region, Madinah, and Qaseem are less likely to smoke. These differences may also reflect differences in the socioeconomic and demographic characteristics of the regions. The findings of this study provide valuable insights into smoking behaviors across different demographics and regions. However, several recommendations can enhance public health interventions, while certain limitations should be acknowledged for future research improvements. One of the most crucial recommendations is the implementation of targeted public health campaigns that address both male and female smokers, particularly among younger age groups (15–24 years), where smoking prevalence is higher. Region-specific awareness programs should also be developed, considering the differences in smoking habits across various regions. Digital platforms, social media, and school-based educational initiatives should be utilized to dissemi-

nate anti-smoking messages effectively. Moreover, stricter legislation and regulatory measures are necessary to curb smoking rates, including higher taxation on tobacco products, more stringent enforcement of smoking bans in public places, and closer monitoring of underage smoking.

Gender-specific interventions should be developed to recognize the distinct smoking patterns between males and females. Smoking cessation support programs tailored for women should be introduced, given the social and cultural stigmas that may prevent them from seeking help. Workplace environments also require urgent attention, as smoking in these spaces remains prevalent. Employers should enforce no-smoking policies, provide cessation support for employees, and promote a smoke-free culture. Educational institutions should integrate anti-smoking education into curricula, targeting students before smoking habits are established. Further research is recommended to investigate the socio-economic determinants of smoking behaviors and assess the impact of alternative nicotine products such as e-cigarettes and vaping. Conducting longitudinal studies would allow researchers to track changes in smoking habits over time and evaluate the effectiveness of anti-smoking policies.

Furthermore, studies should explore the psychological factors influencing smoking initiation and cessation, such as stress, mental health conditions, and peer influence. Despite its valuable contributions, this study has several limitations that should be acknowledged. A significant constraint is the reliance on self-reported data, which may introduce response bias. This is particularly relevant in conservative regions where smoking is stigmatized, leading to underreporting among specific groups, such as women. Furthermore, the study provides a cross-sectional analysis, capturing a

snapshot of smoking prevalence but lacking insights into long-term trends. Future research should adopt a longitudinal approach to better understand smoking initiation, cessation, and relapse rates.

The study also primarily focuses on traditional cigarette smoking, potentially overlooking the growing popularity of alternative nicotine consumption methods, such as shisha (hookah) and vaping. Considering the increasing use of these products, future research should incorporate a broader range of smoking behaviors. Another key limitation is the variability in smoking prevalence across regions, making it difficult to generalize findings to the entire population. Differences in socioeconomic conditions, cultural norms, and policy enforcement levels should be factored into future studies. The study also does not thoroughly analyze the impact of secondhand smoke exposure, which remains a public health concern, particularly in households and workplaces. Additionally, the study may have underrepresented specific demographics, such as expatriates, lower-income individuals, and rural populations, leading to potential gaps in understanding smoking behaviors among these groups. Finally, while the study provides valuable odds, ratios, and likelihood estimates, these do not establish causality. A more robust statistical approach, like regression modeling, could help isolate independent factors influencing smoking behaviors. Furthermore, the data set may not fully reflect recent government-led policies and cultural shifts that affect smoking habits. Future research should include a comparative analysis to assess the impact of these interventions over time.

Chapter 4 also provides a comparative analysis of various 0 and 1 inflated hurdle distributions across synthetic and real-world datasets that demonstrate the superior

performance of the UUTWHurdle model. This model consistently yielded the lowest values for key information criteria such as AIC, BIC, and AICc, as well as the highest log-likelihood values, indicating a better fit to the data. The results were consistent across all three synthetic datasets, each designed with different proportions of zeros, ones, and fractional values, and were further validated using the age dataset. These findings highlight the robustness and flexibility of the UUTWHurdle distribution in modeling data with boundary inflation, making it a reliable choice for similar applications in real-world contexts.

The thesis makes a substantial contribution to public health research by showing how sophisticated statistical models can be used to examine Saudi Arabia's exposure to passive smoking. Policymakers and public health professionals can learn more from the findings, especially highlighting the need for focused interventions in high-risk settings such as workplaces and public areas. Furthermore, the study extends to the literature on GLMs and hurdle models, with multinomial log-linear models, by demonstrating how well they can handle complicated datasets with categorical dependencies and an overabundance of zeros.

5.2 Future Work

Future research should prioritize longitudinal data over cross-sectional studies to assess causal links and the long-term impacts of interventions on passive smoking exposure. More complicated models, such as zero-inflated negative binomial models, should be developed to address zero inflation and overdispersion while increasing analytical pre-

cision and reliability. By controlling for confounding variables, integrating additional covariates such as occupation, education, and socioeconomic status would provide a more comprehensive understanding of the factors contributing to passive smoking. By applying these models to other public health challenges, such as air pollution, we can increase their influence and verify their generalization ability. Geospatial analysis could identify high-risk locations and allow for targeted interventions by mapping exposure patterns, whereas merging machine learning with classical models could enhance predictive accuracy and reveal complicated non-linear interactions. Evaluating the impact of tobacco control initiatives, such as smoking bans or taxes, would provide information for future public health strategies. Investigating cultural and behavioral aspects across demographic groups would allow treatments tailored to individual populations, ensuring cultural appropriateness and effectiveness. Linking exposure data to health outcomes will clarify long-term effects, such as respiratory or cardiovascular problems, thus increasing the evidence base for therapies. International comparisons would provide meaningful insights into global exposure patterns and tobacco control efforts. Addressing these areas in future research would build on the findings of this thesis, advancing the understanding of passive smoking exposure and leading to more accurate predictions, targeted interventions, and effective public health strategies, eventually reducing the global burden of passive smoking.

APPENDIX A

PROOFS

APPENDIX B

SUPPLEMENTARY MATERIAL

REFERENCES

- [1] P. Royston, “Generalized linear models: revision of glm,” *Stata Technical Bulletin*, vol. 3, no. 18, 1994.
- [2] F. J. Carvalho *et al.*, “Análise de experimentos de germinação usando os modelos lineares generalizados,” 2016.
- [3] G. Cerda, C. Pérez, J. I. Navarro, M. Aguilar, J. A. Casas, and E. Aragón, “Explanatory model of emotional-cognitive variables in school mathematics performance: a longitudinal study in primary school,” *Frontiers in Psychology*, vol. 6, p. 1363, 2015.
- [4] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2013.
- [5] J. B. Asimwe, S. Katende, and C. Ashaba, “Impact of non-farm enterprises on household poverty status: A propensity score matching approach,” *Journal of Poverty*, pp. 1–25, 2025.
- [6] J. Mullahy, “Specification and testing of some modified count data models,” *Journal of econometrics*, vol. 33, no. 3, pp. 341–365, 1986.

- [7] Y. Min and A. Agresti, “Random effect models for repeated measures of zero-inflated count data,” *Statistical modelling*, vol. 5, no. 1, pp. 1–19, 2005.
- [8] K. K. Yau, K. Wang, and A. H. Lee, “Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros,” *Biometrical Journal: journal of mathematical methods in biosciences*, vol. 45, no. 4, pp. 437–452, 2003.
- [9] M. G. Chipeta, B. M. Ngwira, C. Simoonga, and L. N. Kazembe, “Zero adjusted models with applications to analysing helminths count data,” *BMC research notes*, vol. 7, pp. 1–11, 2014.
- [10] A. C. Cameron and P. K. Trivedi, “Essentials of count data regression,” *A companion to theoretical econometrics*, vol. 331, 2001.
- [11] C. E. Rose, S. W. Martin, K. A. Wannemuehler, and B. D. Plikaytis, “On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data,” *Journal of biopharmaceutical statistics*, vol. 16, no. 4, pp. 463–481, 2006.
- [12] C. J. Zorn, “Evaluating zero-inflated and hurdle poisson specifications,” *Midwest Political Science Association*, vol. 18, no. 20, pp. 1–16, 1996.
- [13] A. Zeileis, C. Kleiber, and S. Jackman, “Regression models for count data in r,” *Journal of statistical software*, vol. 27, pp. 1–25, 2008.
- [14] A. M. Qattan, M. K. Boachie, M. Immurana, and M. K. Al-Hanawi, “Socioeconomic determinants of smoking in the kingdom of saudi arabia,” *International*

- Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 5665, 2021.
- [15] A. M. Algabbani, R. Almubark, N. Althumiri, A. Alqahtani, and N. BinDhim, "The prevalence of cigarette smoking in saudi arabia in 2018," *Food and Drug Regulatory Science Journal*, vol. 1, no. 1, pp. 1–1, 2018.
- [16] E. C. Koop, C. E. Pearson, and R. M. Schwarz, "Critical issues in global health," *The Journal for Healthcare Quality (JHQ)*, vol. 24, no. 3, pp. 47–51, 2002.
- [17] L. Marshall, M. Schooley, H. Ryan, P. Cox, A. Easton, C. Heaton, K. Jackson, K. C. Davis, G. Homsy, C. for Disease Control, Prevention *et al.*, "Youth tobacco surveillance–united states, 2001-2002," *MMWr surveill summ*, vol. 55, no. 3, pp. 1–56, 2006.
- [18] M. Al-Mana and A. Al-Othaimin, "The unification of the kingdom of saudi arabia," 1982.
- [19] F. Q. Alenzi, *Molecular Targets and Cancer Therapeutics (Part 1)*. Bentham Science Publishers, 2023.
- [20] M. H. Dehghani, P. Bashardoust, D. Nayeri, M. R. Ghalhari, N. B. Yazdi, F. Jajarmi, R. R. Karri, and N. M. Mubarak, "A comprehensive review of the potential outcomes of exposure to tobacco smoke or secondhand smoke," *Health Effects of Indoor Air Pollution*, pp. 167–189, 2024.

- [21] T. Yao, H.-Y. Sung, Y. Wang, J. Lightwood, and W. Max, “Healthcare costs attributable to secondhand smoke exposure at home for us adults,” *Preventive medicine*, vol. 108, pp. 41–46, 2018.
- [22] M. Moradi-Lakeh, C. El Bcheraoui, M. Tuffaha, F. Daoud, M. Al Saeedi, M. Basulaiman, Z. A. Memish, M. A. AlMazroa, A. A. Al Rabeeah, and A. H. Mokdad, “Tobacco consumption in the kingdom of saudi arabia, 2013: findings from a national survey,” *BMC public health*, vol. 15, pp. 1–10, 2015.
- [23] W. H. Organization *et al.*, “Tobacco. fact sheet no 339 [internet]. geneva: Who; 2015 [cited 2016 jan 5]. available from: file,” *C:/Documents% 20and% 20Settings/2509501/Meus% 20documentos/Downloads/WHO% 20Tobacco% 20Fact% 20Sheet. pdf*.
- [24] R. Itumalla and B. Aldhmadi, “Combating tobacco use in saudi arabia: a review of recent initiatives.” *Eastern Mediterranean Health Journal*, vol. 26, no. 7, 2020.
- [25] D. Sengupta and S. R. Jammalamadaka, *Linear models: an integrated approach*. World Scientific, 2003.
- [26] L. Fahrmeir, G. Tutz, W. Hennevogl, and E. Salem, *Multivariate statistical modelling based on generalized linear models*. Springer, 1994, vol. 8.
- [27] A. Dobson and A. Barnett, “An introduction to generalized linear models, set,” *Texts in Statistical Science Series. Tay lor & Francis Group*, 2002.

- [28] M. Goldburd, A. Khare, D. Tevet, and D. Guller, “Generalized linear models for insurance rating,” *Casualty Actuarial Society, CAS Monographs Series*, vol. 5, p. 2016, 2016.
- [29] J. P. Hoffmann, “Generalized linear models: An applied approach,” (*No Title*), 2004.
- [30] N. E. Breslow, “Generalized linear models: checking assumptions and strengthening conclusions,” *Statistica applicata*, vol. 8, no. 1, pp. 23–41, 1996.
- [31] J. Fox, *Applied regression analysis and generalized linear models*. Sage publications, 2015.
- [32] M. Abdulkabir, U. Edem, R. Tunde, and B. Kemi, “An empirical study of generalized linear model for count data,” *Journal of Applied & Computational Mathematics*, vol. 4, no. 5, 2015.
- [33] C. J. Anderson, J. Verkuilen, and T. Johnson, “Applied generalized linear mixed models: Continuous and discrete data,” *Soc Behav Sci*, vol. 63, 2010.
- [34] G. Casella and R. L. Berger, “Statistical inference, thomson learning,” *Duxbury Press: Belmont, CA*, 2002.
- [35] J. Kingman, “Poisson processes oxford univ,” 1993.
- [36] K. Teerapabolarn and K. Jaoun, “Approximation of binomial distribution by an improved poisson distribution,” *International Journal of Pure and Applied Mathematics*, vol. 97, no. 4, pp. 491–495, 2014.

- [37] R. Christensen, *Log-linear models and logistic regression*. Springer Science & Business Media, 2006.
- [38] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media, 2007.
- [39] R. Christensen, “Log-linear models: Springer texts in statistics,” *Springer-Verlag, New York Cornu D & Geoffrion C (1990) Aspects de l’embryogen~ se somatique chez le mrl~ ze. Bull. Soc. Bot. Fr*, vol. 137, pp. 25–34, 1990.
- [40] R. D. Guntur, “The log linear models for two dimensional contingency tables under the multinomial sampling design,” in *FORUM STATISTIKA DAN KOMPUTASI*, vol. 17, no. 2, 2012.
- [41] A. K. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications*. CRC press, 2004.
- [42] Y. Shi, “Advances in big data analytics,” *Adv Big Data Anal*, vol. 10, pp. 978–981, 2022.
- [43] D. L. Olson, Y. Shi, and Y. Shi, *Introduction to business data mining*. McGraw-Hill/Irwin New York, 2007, vol. 10.
- [44] Y. Shi, Y. Tian, G. Kou, Y. Peng, and J. Li, *Optimization based data mining: theory and applications*. Springer Science & Business Media, 2011.
- [45] J. M. Tien, “Internet of things, real-time decision making, and artificial intelligence,” *Annals of Data Science*, vol. 4, pp. 149–178, 2017.

- [46] I. S. Ahmad, R. Jan, P. Nirwan, and P. B. Ahmad, “A new class of distribution over bounded support and its associated regression model,” *Annals of Data Science*, vol. 11, no. 2, pp. 549–569, 2024.
- [47] R. Dumonceaux and C. E. Antle, “Discrimination between the log-normal and the weibull distributions,” *Technometrics*, vol. 15, no. 4, pp. 923–926, 1973.
- [48] P. Embrechts, S. I. Resnick, and G. Samorodnitsky, “Extreme value theory as a risk management tool,” *North American Actuarial Journal*, vol. 3, no. 2, pp. 30–41, 1999.
- [49] R. Kieschnick and B. D. McCullough, “Regression analysis of variates observed on $(0, 1)$: percentages, proportions and fractions,” *Statistical modelling*, vol. 3, no. 3, pp. 193–213, 2003.
- [50] D. O. Cook, R. Kieschnick, and B. D. McCullough, “Regression analysis of proportions in finance with self selection,” *Journal of empirical finance*, vol. 15, no. 5, pp. 860–867, 2008.
- [51] M. Smithson and J. Verkuilen, “A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables,” *Psychological methods*, vol. 11, no. 1, p. 54, 2006.
- [52] M. Ibáñez, M. Prades, and A. Simó, “Modelling municipal waste separation rates using generalized linear models and beta regression,” *Resources, conservation and recycling*, vol. 55, no. 12, pp. 1129–1138, 2011.

- [53] P. Kumaraswamy, “A generalized probability density function for double-bounded random processes,” *Journal of hydrology*, vol. 46, no. 1-2, pp. 79–88, 1980.
- [54] M. Jones, “Kumaraswamy’s distribution: A beta-type distribution with some tractability advantages,” *Statistical methodology*, vol. 6, no. 1, pp. 70–81, 2009.
- [55] I. E. Okorie and E. Afuecheta, “An alternative to the beta regression model with applications to oecd employment and cancer data,” *Annals of Data Science*, vol. 11, no. 3, pp. 887–908, 2024.
- [56] P. R. Tadikamalla and N. L. Johnson, “Systems of frequency curves generated by transformations of logistic variables,” *Biometrika*, vol. 69, no. 2, pp. 461–465, 1982.
- [57] C. W. Topp and F. C. Leone, “A family of j-shaped frequency functions,” *Journal of the American Statistical Association*, vol. 50, no. 269, pp. 209–219, 1955.
- [58] N. Johnson, “Systems of frequency curves derived from the first law of laplace,” *Trabajos de estadística*, vol. 5, no. 3, pp. 283–291, 1955.
- [59] N. L. Johnson, “Systems of frequency curves generated by methods of translation,” *Biometrika*, vol. 36, no. 1/2, pp. 149–176, 1949.
- [60] A. F. B. Menezes, J. Mazucheli, and S. Dey, “The unit-logistic distribution: Different methods of estimation,” *Pesquisa Operacional*, vol. 38, pp. 555–578, 2018.

- [61] J. Mazucheli, A. F. Menezes, and S. Dey, “The unit-birnbaum-saunders distribution with applications,” *Chilean Journal of Statistics*, vol. 9, no. 1, pp. 47–57, 2018.
- [62] J. Mazucheli, A. F. B. Menezes, and S. Chakraborty, “On the one parameter unit-lindley distribution and its associated regression model for proportion data,” *Journal of Applied Statistics*, vol. 46, no. 4, pp. 700–714, 2019.
- [63] M. Ghitany, J. Mazucheli, A. Menezes, and F. Alqallaf, “The unit-inverse gaussian distribution: A new alternative to two-parameter distributions on the unit interval,” *Communications in Statistics-Theory and methods*, vol. 48, no. 14, pp. 3423–3438, 2019.
- [64] R. A. Bantan, C. Chesneau, F. Jamal, M. Elgarhy, M. H. Tahir, A. Ali, M. Zubair, and S. Anam, “Some new facts about the unit-rayleigh distribution with applications,” *Mathematics*, vol. 8, no. 11, p. 1954, 2020.
- [65] B. C. Arnold and R. A. Groeneveld, “Some properties of the arcsine distribution,” *Journal of the American Statistical Association*, vol. 75, no. 369, pp. 173–175, 1980.
- [66] O. E. Barndorff-Nielsen and B. Jørgensen, “Some parametric models on the simplex,” *Journal of multivariate analysis*, vol. 39, no. 1, pp. 106–116, 1991.
- [67] J. R. Van Dorp and S. Kotz, “Modeling income distributions using elevated distributions on a bounded domain,” in *Distribution models theory*. World Scientific, 2006, pp. 1–25.

- [68] G. Moutinho Cordeiro and R. dos Santos Brito, “The beta power distribution,” 2012.
- [69] A. J. Lemonte, W. Barreto-Souza, and G. M. Cordeiro, “The exponentiated kumaraswamy distribution and its log-transform,” 2013.
- [70] F. Castellares and A. J. Lemonte, “On the marshall–olkin extended distributions,” *Communications in Statistics-Theory and Methods*, vol. 45, no. 15, pp. 4537–4555, 2016.
- [71] J. Mazucheli, A. F. Menezes, and S. Dey, “Unit-gompertz distribution with applications,” *Statistica*, vol. 79, no. 1, pp. 25–43, 2019.
- [72] E. Altun and G. M. Cordeiro, “The unit-improved second-degree lindley distribution: inference and regression modeling,” *Computational Statistics*, vol. 35, pp. 259–279, 2020.
- [73] M. Ç. Korkmaz and C. Chesneau, “On the unit burr-xii distribution with the quantile regression modeling and applications,” *Computational and Applied Mathematics*, vol. 40, no. 1, p. 29, 2021.
- [74] M. Ç. Korkmaz and Z. S. Korkmaz, “The unit log–log distribution: A new unit distribution with alternative quantile regression modeling and educational measurements applications,” *Journal of Applied Statistics*, vol. 50, no. 4, pp. 889–908, 2023.
- [75] R. Gupta and D. Kundu, “A new class of weighted exponential distributions,” *Statistics*, vol. 43, no. 6, pp. 621–634, 2009.

- [76] A. Grassia, “On a family of distributions, the argument between 0 and 1 was obtained by transforming the gamma and derived compound distributions,” *Austr J Stat*, vol. 19, no. 2, pp. 108–114, 1977.
- [77] E. Gómez-Déniz, M. Sordo, and E. Calderín-Ojeda, “The log–lindley distribution as an alternative to the beta regression model with applications in insurance,” *Insurance: Mathematics and Economics*, vol. 54, pp. 49–57, 2014.
- [78] A. Bemmaor and N. Glady, “Modeling purchasing behavior with sudden “death”: a flexible customer lifetime model,” *Manage Sci*, vol. 58, no. 5, pp. 1012–1021, 2012.
- [79] R. Bantan, F. Jamal, C. Chesneau, and M. Elgarhy, “Theory and applications of the unit gamma/gompertz distribution,” *Mathematics*, vol. 9, no. 16, 2021.
- [80] S. Gündüz and M. Korkmaz, “A new unit distribution based on the unbounded johnson distribution rule: the unit johnson su distribution,” 2020.
- [81] Y. Benites, V. Cancho, E. Ortega, R. Vila, and G. Cordeiro, “A new regression model on the unit interval: properties, estimation, and application,” *Mathematics*, vol. 10, no. 17, p. 3198, 2022.
- [82] S. Karuppusamy, V. Balakrishnan, and K. Sadasivan, “Improved second-degree lindley distribution and its applications,” *IOSR J Math*, vol. 13, no. 6, pp. 1–10, 2018.

- [83] K. Shekhawat and V. Sharma, "An extension of j-shaped distribution with application to tissue damage proportions in blood," *Sankhya B*, vol. 83, no. 2, pp. 548–574, 2021.
- [84] B. Gompertz, "On the nature of the function expressive of the law of human mortality and on a new mode of determining the value of life contingencies," *Philosophical Transactions of the Royal Society of London*, vol. 115, pp. 513–583, 1825.
- [85] W. Willemse and H. Koppelaar, "Knowledge elicitation of gompertz law of mortality," *Scandinavian Actuarial Journal*, vol. 2, pp. 168–179, 2000.
- [86] R. Makany, "Theoretical basis of gompertz's curve," *Biometrical Journal*, vol. 33, pp. 121–128, 1991.
- [87] P. Franses, "Fitting a gompertz curve," *Journal of the Operational Research Society*, vol. 45, pp. 109–113, 1994.
- [88] Z. Chen, "Parameter estimation of the gompertz population," *Biometrical Journal*, vol. 39, pp. 117–124, 1997.
- [89] A. El-Gohary, A. Alshamrani, and A. Al-Otaibi, "The generalized gompertz distribution," *Applied Mathematical Modelling*, vol. 37, pp. 13–24, 2013.
- [90] M. Khan, K. Robert, and L. Irene, "Transmuted gompertz distribution: Properties and estimation," *Pakistan Journal of Statistics*, vol. 32, no. 3, pp. 161–182, 2016. [Online]. Available: <http://hdl.handle.net/1959.13/1331685>

- [91] A. El-Bassiouny, M. El-Damcese, A. Mustafa, and M. Eliwa, “Mixture of exponentiated generalized weibull-gompertz distribution and its applications in reliability,” *Journal of Statistics Applications and Probability*, vol. 5, no. 3, pp. 455–468, 2016.
- [92] ———, “Exponentiated generalized weibull-gompertz distribution with application in survival analysis,” *Journal of Statistics Applications and Probability*, vol. 6, no. 1, pp. 7–16, 2017.
- [93] A. Drapella, “Complementary weibull distribution: Unknown or just forgotten,” *Quality and Reliability Engineering International*, vol. 9, no. 4, pp. 383–385, 1993.
- [94] R. Jiang, M. Zuo, and H. Li, “Weibull and inverse weibull mixture models allow negative weights,” *Reliability Engineering and Systems Safety*, vol. 66, pp. 227–234, 1999.
- [95] A. El-Gohary, A. El-Bassiouny, and M. El-Morshedy, “Inverse flexible weibull extension distribution,” *International Journal of Computer Applications*, vol. 115, no. 2, pp. 46–51, 2015.
- [96] M. El-Morshedy, A. El-Bassiouny, and A. El-Gohary, “Exponentiated inverse flexible weibull extension distribution,” *Journal of Statistics Applications and Probability*, vol. 6, no. 1, pp. 169–183, 2017.

- [97] M. Eliwa, M. El-Morshedy, and M. Ibrahim, “Inverse gompertz distribution: properties and different estimation methods applied to complete and censored data,” *Annals of Data Science*, vol. 6, no. 2, pp. 321–339, 2019.
- [98] R. Alshenawy, “A new one parameter distribution: properties and estimation with applications to complete and type ii censored data,” *Journal of Taibah University for Science*, vol. 14, no. 1, pp. 11–18, 2020.
- [99] A. Marshall and I. Olkin, “A new method for adding a parameter to a family of distributions with application to the exponential and weibull families,” *Biometrika*, vol. 84, pp. 641–652, 1997.
- [100] K. Zografos and N. Balakrishnan, “On families of beta- and generalized gamma-generated distributions and associated inference,” *Statistical Methodology*, vol. 6, pp. 344–362, 2009.
- [101] G. Cordeiro and M. de Castro, “A new family of generalized distributions,” *Journal of Statistical Computation and Simulation*, vol. 81, no. 7, pp. 883–898, 2011.
- [102] C. Alexander, G. Cordeiro, E. Ortega, and J. Sarabia, “Generalized beta generated distributions,” *Computational Statistics and Data Analysis*, vol. 56, pp. 1880–1897, 2012.
- [103] M. Ristic and N. Balakrishnan, “The gamma-exponentiated exponential distribution,” *Journal of Statistical Computation and Simulation*, vol. 82, no. 8, pp. 1191–1206, 2012.

- [104] G. Cordeiro, E. Ortega, and D. da Cunha, “The exponentiated generalized class of distributions,” *Journal of Data Science*, vol. 11, pp. 1–27, 2013.
- [105] M. Bourguignon, R. Silva, and G. Cordeiro, “The weibull-g family of probability distributions,” *Journal of Data Science*, vol. 12, pp. 53–68, 2014.
- [106] M. Tahir, G. Cordeiro, A. Alzaatreh, M. Mansoor, and M. Zubair, “The logistic-x family of distributions and its applications,” *Communications in Statistics-Theory and Methods*, vol. 45, no. 24, pp. 7326–7349, 2016.
- [107] A. El-Bassiouny and M. El-Morshedy, “The univariate and multivariate generalized slash student distribution,” *International Journal of Mathematics and Its Applications*, vol. 3, no. 3-B, pp. 35–47, 2015.
- [108] R. ZeinEldin, S. Hashmi, M. Elsehety, and M. Elgarhy, “Type ii half logistic kumaraswamy distribution with applications,” *Journal of Function Spaces*, 2020.
- [109] G. Zuo, K. Fu, X. Dai, and L. Zhang, “Generalized poisson hurdle model for count data and its application in ear disease,” *Entropy*, vol. 23, no. 9, p. 1206, 2021.
- [110] R. Ospina and S. Ferrari, “Inflated beta distributions,” *Statistical Papers*, vol. 51, pp. 111–126, 2010.
- [111] F. Cribari-Neto and J. Santos, “Inflated kumaraswamy distributions,” *Anais da Academia Brasileira de Ciências*, vol. 91, 2019.

- [112] G. Martínez-Flórez, R. Tovar-Falón, and C. Barrera-Causil, “Inflated unit-birnbaum-saunders distribution,” *Mathematics*, vol. 10, no. 4, p. 667, 2022.
- [113] R. K. Iyer, “Hazard and reliability functions; failure rate,” 2013, unpublished or informal reference.
- [114] N. Kantas, A. Doucet, S. Singh, J. Maciejowski, and N. Chopin, “On particle methods for parameter estimation in state-space models,” 2015.
- [115] J. Hosking, “L-moments: Analysis and estimation of distributions using linear combinations of order statistics,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 52, no. 1, pp. 105–124, 1990.
- [116] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä, “A survey of monte carlo methods for parameter estimation,” *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 1, pp. 1–62, 2020.
- [117] E. Hooda, B. Hooda, and N. Tanwar, “Probability weighted moments (pwms) and partial probability weighted moments (ppwms) of type-ii extreme value distribution,” 2020.
- [118] I. Myung, “Tutorial on maximum likelihood estimation,” *Journal of Mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.
- [119] F. Richards, “A method of maximum-likelihood estimation,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 23, no. 2, pp. 469–475, 1961.

- [120] A. Henningsen and O. Toomet, “maxlik: A package for maximum likelihood estimation in r,” *Computational Statistics*, vol. 26, pp. 443–458, 2011.
- [121] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: <https://www.R-project.org/>
- [122] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [123] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, vol. 6, pp. 461–464, 1978.
- [124] C. Hurvich and C. Tsai, “Regression and time series model selection in small samples,” *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.
- [125] N. Balakrishna and K. Shiji, “On a class of bivariate exponential distributions,” *Statistics and Probability Letters*, vol. 85, pp. 153–160, 2014.
- [126] K. Okhli and M. Nooghabi, “On the contaminated exponential distribution: a theoretical bayesian approach for modeling positive-valued insurance claim data with outliers,” *Applied Mathematics and Computation*, vol. 392, p. 125712, 2021.
- [127] S. Chakraborty and S. Bhattacharjee, “Modeling proportion of success in high school leaving certificate examination- a comparative study of inflated unit lindley and inflated beta distribution,” *Journal of Mathematics and Computer Science*, vol. 11, no. 6, pp. 7170–7198, 2021.

- [128] P. Zhang, Z. Qiu, and C. Shi, “Simplexreg: An r package for regression analysis of proportional data using the simplex distribution,” *Journal of Statistical Software*, vol. 71, no. 11, 2016.

VITAE

- Name: NAFISA MAHMUD NASIR
- Nationality: NIGERIAN
- Date of Birth: 09 FEBRUARY, 1993
- Email: *nafnas069@gmail.com*
- Permanent Address: No. 35 RIKKOS NEWLAYOUT JOS,
PLATEAU STATE, NIGERIA

Education

- [2025:] MSc in Statistics, King Fahd University of Petroleum and Minerals (KFUPM)
- [2022:] Credit, Teachers Registration Council of Nigeria
- [2021:] Postgraduate Diploma in Education (PGDE), National Teacher's Institute, Jos, Nigeria

- [2020:] National Youth Service Corps (NYSC), Bright Future Academy, Bauchi State, Nigeria
- [2019:] BSc in Statistics, University of Jos, Second Class Upper Division, CGPA: 4.33

Experience

- [2023–2025:] Lab Recitation Instructor, Mathematics Department, KFUPM
Teaching Python, MATLAB, and R for Math 101, 102, and STAT 319.
- [2020–2023:] Teacher, Al Hilal Secondary School, Jos.
- [2020–2023:] Teacher, Al Iklas Secondary School, Jos.
- [2015–2019:] Teacher, Cambridge private School, Jos.
- [2011–2014:] Teacher, voluntary in Usama, Maljausunnah Private school, and others, Jos.

Skills

- Proficient in R, Python, MATLAB, LaTeX, Minitab, and Microsoft Word.
- Strong interpersonal, communication, and problem-solving skills.

- Experience with data analysis and statistical modeling.
- Ability to work independently with minimal supervision.

References

- Dr. Emmanuel Osita Afuecheta
Assistant Professor, King Fahd University of Petroleum and Minerals
emmanuel.afuecheta@kfupm.edu.sa
- Dr. Jesus Igor Heberto Barahona Torres
Visiting Assistant Professor, King Fahd University of Petroleum and Minerals
j.barahonatorres@kfupm.edu.sa