KADİR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES
PROGRAM OF COMPUTER ENGINEERING

INVESTIGATION THE RISK OF AUTISM BY EVALUATING PRENATAL AND POSTNATAL EXPOSURE TO TRAFFIC-RELATED AIR POLLUTION

TAMER DEMİR

MASTER’S THESIS

ISTANBUL, JUNE, 2020

Tamer DEMİR M.S. Thesis 2020

INVESTIGATION THE RISK OF AUTISM BY EVALUATING PRENATAL AND POSTNATAL EXPOSURE TO TRAFFIC-RELATED AIR POLLUTION

TAMER DEMİR

MASTER’S THESIS

Submitted to the School of Graduate Studies of Kadir Has University in partial fulfillment of the requirements for the degree of Master’s in the Program of Computer Engineering

ISTANBUL, JUNE, 2020

DECLARATION OF RESEARCH ETHICS / METHODS OF DISSEMINATION

I, TAMER DEMİR, hereby declare that;

- this Master's Thesis is my own original work and that due references have been appropriately provided on all supporting literature and resources;
- this Master's Thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- I have followed "Kadir Has University Academic Ethics Principles" prepared in accordance with the "The Council of Higher Education's Ethical Conduct Principles"

In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.

Furthermore, both printed and electronic copies of my work will be kept in Kadir Has Information Center under the following condition as indicated below:

□ The full content of my thesis/project will be accessible from everywhere by all means.

TAMER DEMİR
24/06/2020

KADIR HAS UNIVERSITY

SCHOOL OF GRADUATE STUDIES

ACCEPTANCE AND APPROVAL

This work entitled **INVESTIGATION THE RISK OF AUTISM BY EVALUATING PRENATAL AND POSTNATAL EXPOSURE TO TRAFFIC-RELATED AIR POLLUTION** prepared by TAMER DEMİR has been judged to be successful at the defense exam held on **24/06/2020** and accepted by our jury as **TYPE OF THE THESIS**.

APPROVED BY:

Assoc. Prof. Dr. Tamer DAĞ (Advisor) Kadir Has University _______________

Assoc. Prof. Dr. Habib ŞENOL Kadir Has University _______________

Assoc. Prof. Dr. Tansal GÜÇLÜOĞLU Yıldız Technical University _______________

I certify that the above signatures belong to the faculty members named above.

____

Prof. Dr. Sinem AKGÜL AÇIKMEŞE

Dean of School of Graduate Studies

DATE OF APPROVAL: 24/06/2020

TABLE OF CONTENTS

INVESTIGATION THE RISK OF AUTISM BY EVALUATING PRENATAL AND POSTNATAL EXPOSURE TO TRAFFIC-RELATED AIR POLLUTION

ABSTRACT

Autism spectrum disorder (ASD) which is a group of neurodevelopmental disorder that appears during the first few years of a child's life affecting a child's communication and socialization abilities with increasing prevalence. Recently, several recent studies have found associations between exposure to traffic-related air pollution (TRAP) and ASD. The primary aim of this study is to investigate/examine the relation between TRAP and four air pollutants (NO_2, O_3, PM_{10}, $PM_{2.5}$) and ASD during prenatal or post-natal by using multiple logistic regression models and variable selection methods. Results show that the adjusted odds ratio (AOR) for ASD per IQR increase was strongly associated for exposure to NO_2 during the first year period, was moderately associated for exposure to NO_2 (from interstate highways during the third trimester; from the county highway during the first year; from city street during the first year; from all roads during the all pregnancy; from all roads during the first trimester) and O_3 during the second year, and weakly associated with exposure to NO_2 from interstate highways during the second trimester, O_3 during the first trimester and $PM_{2.5}$ during the second year. Additionally, comparing fourth to first quartile exposures the AOR was 15.47 for NO_2 from interstate highways during the third trimester, was 5.00 for NO_2 from all roads during the first trimester, and comparing third to first quartile exposures the AOR was 2.31 for $PM_{2.5}$ during the second year. As a result, a strong relationship between NO_2 exposure and ASD was detected for each 7.1 ppb [IQR] increase in NO_2 during the first year and subjects exposed to a higher level of NO_2 during the first and third trimester, and $PM_{2.5}$ during the second year was also associated with increased risk of ASD.

Keywords: Autism spectrum disorder (ASD), Air pollution, Multiple logistic regression, and Variable selection

DOĞUM ÖNCESİ VE SONRASI TRAFİK KAYNAKLI HAVA KİRLİLİĞİNE MARUZ KALMA İLE OTİZM SPEKTRUM BOZUKLUĞU (ASD) ARASINDAKİ İLİŞKİYİ ORTAYA ÇIKARMAK

ÖZET

Otizm spektrum bozukluğu (ASD), hızlıca artan bir yaygınlıkla çocukluğun ilk yıllarında ortaya çıkan ve çocuğun iletişim ve sosyalleşme yetilerini etkileyen nörogelişimsel bozuklular grubudur. Son zamanlarda trafik kaynaklı hava kirliliğine (TRAP) maruz kalma ve Otizm spektrum bozukluğu (ASD) arasındaki ilişkiyi ortaya çıkaran çalışmalar yapılmıştır. Bu çalışmanın temel amacı, trafik kaynaklı hava kirliliği ve 4 hava kirleticisi (NO_2, O_3, PM_{10}, $PM_{2.5}$) ve Otizm spektrum bozukluğu (ASD) arasındaki ilişkiyi lojistik regresyon modellerini ve değişken seçim yöntemlerini kullanarak ortaya çıkarmak / incelemektir. Sonuçlar, [IQR] artışı başına ASD için ayarlanmış olasılık oranının (AOR), ilk yıl boyunca NO_2'ye maruz kalmayla güçlü, ikinci yıl boyunca O_3'e, NO_2'ye (üçüncü üç aylık dönemde eyaletler arası otoyollardan; ilk yıl boyunca ilçe otoyolundan; ilk yıl boyunca şehir caddesinden; tüm gebelik boyunca tüm yollardan; ilk üç aylık dönemde tüm yollardan çıkan) maruz kalmayla orta derecede, ve ikinci üç aylık dönemde eyaletler arası otoyollardan çıkan NO_2'ye ilk üç aylık dönemde O_3'e ve ikinci yılda $PM_{2.5}$'e maruz kalma ile zayıf bir şekilde ilişkili olduğunu göstermektedir. Ek olarak, dördüncü ve birinci çeyrek maruz kalmaları karşılaştırıldığında, üçüncü üç aylık dönemde eyaletler arası otoyollardan çıkan NO_2 için AOR 15.47, ilk üç aylık dönemde tüm yollardan çıkan NO_2 için 5.00 idi ve üçüncü ile birinci çeyrek maruz kalmaları karşılaştırıldığında, AOR ikinci yılda $PM_{2.5}$ için 2.31 idi. Sonuç olarak, ilk yıl NO_2'deki her 7.1 ppb [IQR] artış için NO_2'ye maruz kalma ile ASD arasında güçlü bir ilişki tespit edildi ve ayrıca birinci ve üçüncü üç aylık dönem boyunca NO_2'nin, ikinci yıl boyunca $PM_{2,5}$'nin daha yüksek düzeyine maruz kalan denekler artan ASD riski ile ilişkilendirildi.

Anahtar Sözcükler: Otizm spektrum bozukluğu (ASD), Hava kirliliği, Çoklu lojistik regresyon

ACKNOWLEDGEMENTS

I would like to express my appreciation to all people who gave a contribution to my thesis. First and foremost, I would like to thank Assoc. Prof. Dr. Tamer Dağ for his support and guidance throughout the thesis process.

I would also like to thank my mum and my wife, Hilal, for the support they provided me throughout my entire life.

DEDICATION

To my wife and mother

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASD	Autism spectrum disorder
CDC	Centers for Disease Control and Prevention
ADDM	The Autism and Developmental Disabilities Monitoring
WHO	World Health Organization
TRAP	Traffic-related air pollution
HAP	Hazardous air pollutant
NDAR	National Database repository for Autism Research
ADHD	Attention deficit hyperactivity disorder
CHD	Coronary Heart Disease
MLE	Maximum Likelihood Estimation
AIC	Akaike Information Criterion
BIC	Bayesian information criterion
LRT	Likelihood Ratio Test
LR	Likelihood Ratio
GOF	Goodness Of Fit
IQR	Interquartile Range
NIH	National Institutes of Health
AOR	Adjusted Odds Ratio
OR	Odds Ratio

1. INTRODUCTION

1.1 Background to the Study

Autism spectrum disorder (ASD) is considered as a group of neurodevelopmental disorder that is characterized by difficulties with communication and social interaction and restricted, repetitive and stereotyped behaviors, interests, and activities present in early childhood [1, 2]. The prevalence of ASD was estimated 1 in 59 in the US according to the CDC's ASD prevalence report [3] in April 2018 and was 4 times more common in boys than girls. It also showed a dramatic rise, approximately from 0.67 (1 in 150) in 2002 to 1.69 (1 in 59) in 2014 [4]). The symptoms are generally evident in the early developmental life of a child, usually 2-3 years old and many children are not diagnosed as soon as the baby is born.

Recently, many researchers have been trying to find out the causes of ASD, but they couldn't find the exact causes of ASD. Epidemiological studies suggest that both genetics and environment likely play a role in ASD [5, 6]. The environmental factors were suggested to account for around 40% for autism [7, 8]. Similarly, the genetic factors are responsible for around 50% of the risk for ASD [8]. So, some studies have demonstrated that understanding the contribution of environmental factors to ASD might be easier than genetics and helpful on the increasing prevalence of ASDs [9].

1.2 Statement of the Problem and Motivation

The recent increase in brain disorders such as ASD, ADHD, and Down syndrome and the birth of my nephew with Down syndrome have motivated me to work on these issues. I asked myself "Why the prevalence of autism has increased without sufficient explanation and what can I do?"

When we think about what is going on parallel in the world with the rise of autism, it can be seen that the environment is excessively polluted by toxic wastes caused by urbanization, industrialization, and increase in the transportation due to globalization, that is, increase in the number of vehicles.

So, environmental factors or exposures including all nongenetic factors, from viruses and medications to chemicals agents during prenatal, natal, and postnatal development may influence brain development, leading to neurodevelopmental abnormalities that can contribute to ASD [7].

Researcher have noted several environmental risk factors related to ASD before and during birth, such as increased maternal and paternal age, maternal health during pregnancy, maternal lifestyle, pregnancy complications and prenatal exposure to environmental toxins(heavy metals, pesticides, industrial pollutants, and air pollution)

As a result, in my study, as a criminal of autism, air toxins that pollute the environment, especially traffic-related pollution (TRAP) made me motivated to study ASD despite the genetic susceptibility.

1.3 Objectives

This study aims to investigate the relation between traffic-related air pollution (TRAP) and four air pollutants (NO_2, O_3, PM_{10}, $PM_{2.5}$) and ASD during prenatal or post-natal

1.4 Methodology

The available data for my research has been collected from the National Database repository for Autism Research (NDAR) data repository [10].

My working data is a collection whose title is “Distance to Freeway and Major Road” and its investigator is Rob McConnell. This collection contains birth address-based measures on the distance to the interstate highway, state highway, major road, and local road for study subjects enrolled in the CHARGE study, and the subjects for this project were pregnant mothers.

Multiple logistic regression was used to test the relation between air pollutants (NO_2, O_3, PM_{10}, $PM_{2.5}$) and ASD.

The likelihood ratio and Wald test were used to illustrate how the significance of regression parameters in multiple logistic regression.

Since the number of explanatory variables is large (i.e., 40 or more), variable selection methods were used to choose the best subset of the predictors among many variables associated with the outcome.

1.5 Significance of the Study

The early years of a child's life particularly the period from birth to 2 years old are very important for brain development. The brain grows incredibly in two years. 80% of brain development is completed in this period. During this stage, children are highly influenced by the environment. Many factors in addition to genes such as inadequate nutrition and exposure to toxins or infections during and after pregnancy affect brain development. This period is likely to be critical in neurodevelopmental disorders including autism [11, 12].

So, the first reason why this study is important for me is that I focused on the air pollutant exposures from four different roads occurring from the first trimester through the child's first year which are the periods of brain development. Secondly, we tried to make clear the role of timing for TRAP during pregnancy and early life and tested the relation between TRAP and ASD. Finally, identifying air pollutant exposures that contribute to autism is important to prevent ASD by reducing ambient pollution with the help of government policies and therefore informing a large number of people especially sensitive groups of pregnant women and children.

1.6 Organization of Study

In this project, the data mining techniques were used according to CRISP-DM, in which a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 1.1. Note that the iterative nature of CRISP is symbolized by the outer circle, the significant dependencies between phases are indicated by the arrows. The phase sequence can be changed due to different conditions. That is, the next phase in the sequence often depends on the outcomes associated with the previous phase [13].

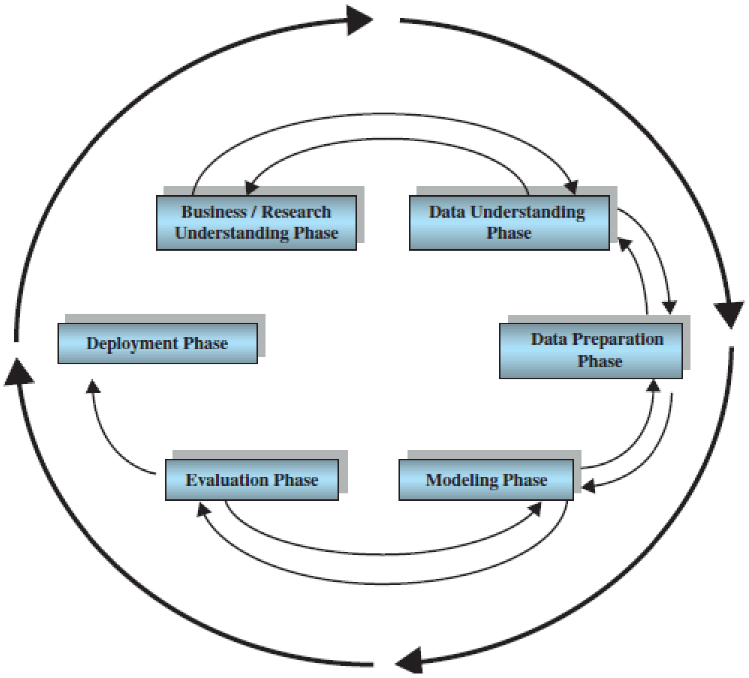


Figure 1.1: Cross-Industry Standard Process for Data Mining (CRISP-DM) [13]

The organization of this thesis as follows:

Chapter 1 deals with the background of the study (ASD and causes), statement of the problem (motivation), objectives of the study, methodology, and significance of the study.

Chapter 2 introduces the ASD domain, what ASD is, its causes and prevalence, and environmental factors playing role in ASD. It also examines the related literature review.

Chapter 3 is a data understanding phase that describes the source of data, datasets, and data fields.

Chapter 4 gives the theoretical background, the applied data preprocessing, logistic regression, and variable selection methods.

Chapters 5 covers the analysis of data and evaluates the quality of the data, cleaning the raw data, and dealing with missing and outlier data before proceeding to the modeling phase.

Chapter 6 covers model building and model evaluation phases. It presents the implementation of logistic regression, how to fit the logistic regression model, how to teste the significance of coefficients, and interpret the model coefficients. Additionally, it describes the variable selection methods and provides the results for each model in the study.

Chapter 7 presents a summary of the main findings achieved in Chapter 6.

Chapter 8 presents the derived conclusion of the study.

2. AUTISM & LITERATURE REVIEW

2.1 What is Autism?

Autism spectrum disorder (ASD) is a group name given to a collection of neurodevelopmental disorders such as Autistic Disorder, Childhood Disintegrative Disorder, Asperger Syndrome, Rett’s Disorder, and Pervasive Developmental Disorder. Its symptoms and characteristics can be described as a lack of social dialog with other people, lack of communication skills, repetitive, stereotypical attitudes, body movements especially appearing in the early developmental life of a child. These symptoms and characteristics can occur in different combinations and degrees ranging from mild to severe and also no two Autistic children resemble each other [14, 10, 15].

2.2 Prevalence of ASD

Autism prevalence shows how common ASD is in the general population. In the United States, an ASD prevalence estimates report [16] was published by the Centers for Disease Control and Prevention (CDC) in 2018. The report subjects’ data whose records of 8 years old children living in 11 areas of the United States were collected by The Autism and Developmental Disabilities Monitoring (ADDM) Network surveillance system during 2014. The summary of the ASD prevalence report is shown in Table 2.1 [17].

Table 2.1: Prevalence of ASD in 8-year-olds (2014) [18].

		Prevalence	
		1 in x	**Prevalence %**
Sex	**Boys**	1 / 38	2.70%
	Girls	1 / 152	0.70%
Race / Ethnicity	**White**	1 / 58	1.70%
	Black	1 / 63	1.60%
	Asian / Pacific Islander	1 / 74	1.40%
	Hispanic	1 / 71	1.40%
Overall		1 / 59	1.70%

According to Table 2.1,

- The overall average prevalence of ASD 1 in every 59
- ASD is 4 times more common in boys than girls. For example, it is 0.7% for girls, whereas 2.7% in boys.
- ASD occurs in the children of different races and ethnic groups and the prevalence for non-Hispanic white children is higher than non-Hispanic black children.

The results of many studies conducted in Asia, Europe, and North America were found to be between 1% and 2% for average prevalence [19]. The estimated prevalence of ASD with the collected data of the children living in 11 different sites between 2000 and 2014 years across the US are shown in Table 2.2 [19]. It is concluded that the estimated prevalence of ASD increased by approximately 16% between 2012 and 2014, 25% between 2006 and 2008, 71% between 2002 and 2008, and 100% between 2004 and 2014.

Table 2.2: Prevalence of ASD based on ADDM Network studies published from 2007 to 2018 (surveillance years 2000-2014) [19].

Surveillance Year	Birth Year	# of ADDM Sites	Combined Prevalence per 1,000 Children	1 in X children
2000	1992	6	6.7	1 / 150
2002	1994	14	6.6	1 / 150
2004	1996	8	8	1 / 125
2006	1998	11	9	1 / 110
2008	2000	14	11.3	1 / 88
2010	2002	11	14.7	1 / 68
2012	2004	11	14.5	1 / 69
2014	2006	11	16.8	1 / 59

Based on the data in Table 2.2 dramatic increases in the prevalences of ASD can be seen in Figure 2.1

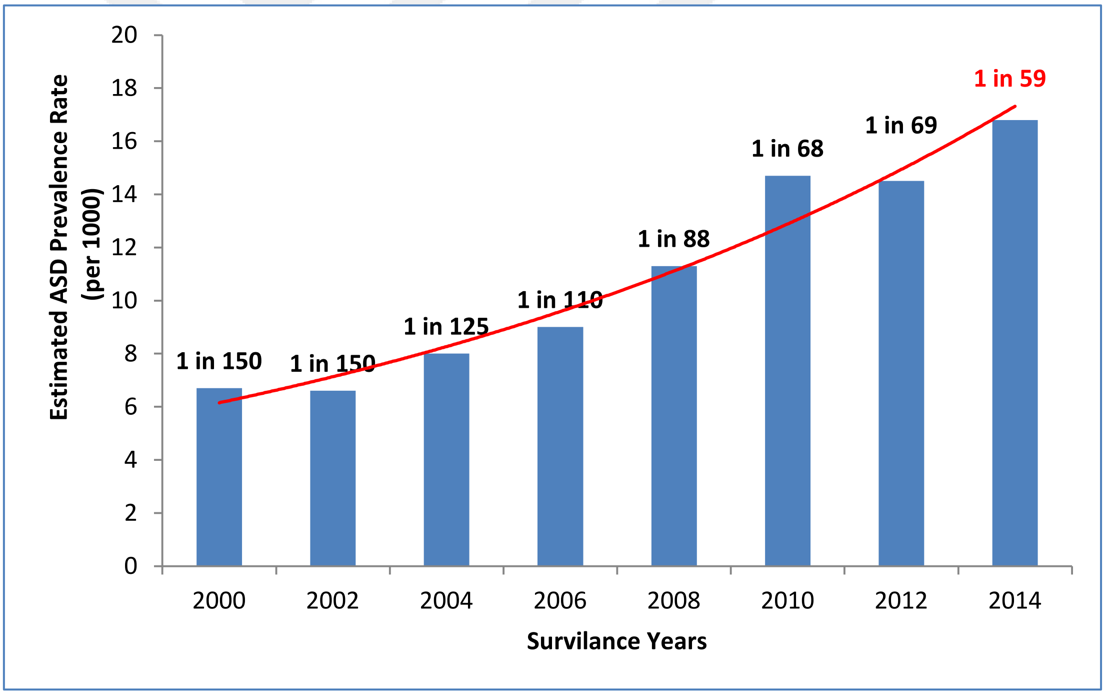


Figure 2.1: The Estimated ASD Prevalence Rate (per 1000) based on Table 2.2 Data

Finally, while ASD was a rare disease, now it is 1 in 59 among children. Why this dramatic growth has been? It may be explained with better diagnosis and more advanced diagnostic criteria [20].

2.3 Causes and Risk Factors

Recently, increasing autism prevalence rates have been found in many studies without enough explanation about it. Although scientists are still trying to understand the causes of ASD, they couldn’t manage to identify the exact causes of autism. There may be many risk factors that can cause a child more likely to have an ASD. It is generally believed that genetics and environmental factors both play a role in ASD [5, 6].

Genetic risk factors are present at birth especially in DNA such as gene mutations, gene deletions, or duplications, however environmental risk factors or all non-genetic factors may cause to develop ASD during prenatal or postnatal periods [21].

2.3.1 Genetic Factors

Many researcher believe that there is a strong role of genes in the development of ASD [22]. Twin and family studies support the genetic theory [23]. For example, ASD is 50 to 200 times more common in siblings of autistic probands than normal population [22] and, parents who have a child with ASD have a 2%–18% chance of a second affected child [24]. While the contribution of genetic factors to ASD was 7–8% of autism cases in 2010 [25], but this contribution has increased about 10-30% with the help of higher diagnostic tools [26].

2.3.2 Environmental Factors

The environment can be defined as a combination of external physical factors that affect the development of a child. These factors can be either viruses, medications or chemicals, and physical agents [27]. In fact, the first environment for a child is the womb in which the baby develops, and it is involved in the development of autism. Environmental factors may influence brain development at different stages. They act together in harmony with susceptible genes. They are some interactions with each other, which may lead to changes in gene expression. Additionally, genes may indirectly change the biochemistry of the brain by affecting the metabolism and activity of foreign chemicals such as pesticides. Changes to DNA, not inherited from parents, causing damage to the genetic code environmental exposures are associated with ASD risk [27]. Scientists believe that de novo mutations lead to some children to susceptible to ASD when exposed to certain environmental factors [28].

Environmental factors are thought to be responsible for around 40% for ASD and 10–40% of the risk for ADHD [29].

Changes in over 1,000 genes have been thought to affect the risk of developing ASD, but most of these variations have only a small effect when combined with environmental risk factors, such as increased maternal age, pregnancy complications, and others that have not been identified. “Non-genetic factors may contribute up to about 40 percent of ASD risk” [30].

The different environmental risk factors may be associated with ASD involves events before and during birth, such as

- Increased maternal and paternal age
- Maternal Health during pregnancy (Maternal stress, maternal obesity, diabetes   or immune factors)
- Maternal Lifestyle (Medication usage such as prenatal vitamins and folic acid and related nutrients, substance use, drug usage, alcohol usage, and smoking)
- Pregnancy complications
    - Metabolic complications like gestational diabetes,
    - Delivery complications such as birth asphyxia leading to oxygen deprivation of the baby’s brain,
    - Neonatal complications such as low birth weight, preterm birth.
- Prenatal or postnatal exposure to environmental toxins
    - Heavy metals such as lead and mercury,
    - Pesticides such as organophosphate pesticides (OPs) or organochlorines pesticides (OCPs),

- Industrial Pollutants like Phthalates or Polychlorinated Biphenyls(PCB),
- Air pollution such as hazardous air pollutants(benzene, methylene chloride), criteria air pollutants (NO_2, O_3, $PM_{2.5}$, PM_{10}), metals(lead, cadmium), and traffic-related pollution (TRAP) [27, 31].

In my study, criteria air pollutants (especially NO_2, O_3, $PM_{2.5}$, PM_{10}), and TRAP were focused due to the working datasets.

2.3.2.1 Air pollution

Air pollution which is also called air toxics or air pollutants is a combination of many kinds of gases, droplets, and solid particles in the air. Because of the increase in population and the need for energy, it has become a very important problem all over the world [32, 33, 34].

Air pollution has been mostly caused by the air toxics emitted from different sources, primarily transportation(e.g., automobile, trucks, buses exhaust), secondly industrial emissions (e.g., factories, refineries, power generation) and thirdly indoor sources(e.g. smoking, cooking, heating, and lighting, vapors from building materials, paints)
[35, 36, 37].

There are many types of air pollutants, but the major ones that make the air quality worse are particulate matter (PM), sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3) and lead [38].

According to the World Health Organization (WHO) statistics reports that around 7 million people die every year from exposure to polluted air [39]. Additionally, hazardous air pollutants (HAPs) have negative impacts on health, they may cause cancer or other serious health effects, such as neurological or respiratory disease and birth or developmental defects[38].

There is increasing evidence that TRAP which is emitted by vehicles especially in urban areas contributes to air quality and may affect pregnancy outcomes and child development [40, 41].
It has been reported that the emissions caused by the transportation sector are approximately 55%, 10%, and 10% for NO_x, VOCs, and $PM_{2.5}$ and PM10 respectively in the U.S [41].

The traffic-related air pollutants (NO_2, NO, O_3, SO2, CO, and PM) were mainly focused on in these studies. Most of the studies showed a positive association between maternal exposure to the abovementioned pollutants and ASD.

In 2011, Volk et al. [1] investigated the relationship between distance to the major roads of the children's homes and ASD. Children who lived within 309m of a freeway have higher odds of having autism than children who lived bigger than 1149 m from a freeway [1]. Further, in 2013, Volk et al. [2] further discovered that exposure to TRAP, NO_2, $PM_{2.5}$, and PM_{10} during the first year of life was associated with autism.

2.3.2.1.1 Nitrogen Oxides (NO_x)

Nitrogen oxides (NO_x) are a group of gases that are composed of nitrogen(N) and oxygen (O). Two of the most common nitrogen oxides are nitric oxide (NO) and nitrogen dioxide (NO_2). Nitrogen oxides are emitted from motor vehicle exhaust and high-temperature combustion processes such as power plants or industrial plants by burning of coal, *oil*, *diesel* fuel, and *natural gas*.

Particulate matter and ground-level ozone are formed by the reaction NO_x with sunlight or other chemicals in the air. Acid rains are also formed by the end of the interaction of NO_x with water, oxygen, and other chemicals(eg., sulfur dioxide) in the atmosphere.

If you live near power or industrial plants or if you are in heavy traffic, further if you smoke cigarettes, you can be exposed to nitrogen oxides by breathing air [42, 43].

There have been some studies finding the relation between ASD and NO_2 in literature.

In 2013, Volk et al. [2] found that exposure to nitrogen dioxide during gestation was also associated with ASD. Then, Ritz et al. [44] found NO_2 exposure in all trimesters to be associated with ASD in 2018. Finally, in 2109, Oudina et al. [4], found that exposure to nitrogen dioxide during the prenatal period was associated with autism. On the other hand, Gong et al. [3], Gong et al. [45], Pagalan et al. [46], Raz et al. [34] studies reported no associations.

2.3.2.1.2 Particulate Matter

Particulate matter (PM) is a mixture of solid or liquid tiny particles suspended in the air which are approximately 1 to 10 micrometers(μm) in size.
The Particulate matter (PM) is identified according to aerodynamic diameter

- **PM_{10}**: inhalable coarse particles with a diameter smaller than 10 μm
- **$PM_{2.5}$**: fine inhalable small particles with a diameter smaller than 2.5 μm

Many particles are emitted into the atmosphere from transportation (vehicles), industry (factory), combustion of fossil fuels, natural (e.g. by dust storms). These particles are also formed in the atmosphere as a result of complex chemical reactions between pollutants such as sulfur dioxide and nitrogen oxide.

Since these particles are very small, they are more dangerous when breath, they may reach inside the lungs. They have important effects on human health, such as cardiovascular, lung, skin diseases, and sometimes cause premature deaths [34, 37, 47, 48].

$PM_{2.5}$ were positively associated with ASD in the following studies:
Becerra et al. [37] and Volk et al. [2] in 2013, Talbott et al. [49] and Raz et al.[39] in 2015, Chen et al. [50] and Ritz et al. [44] in 2018, Geng et al. [51] and Jo et al. [52] in 2019.

However, in 2 other studies, Guxens et al.[53] and Pagalan et al. [54], found no association between ASD and $PM_{2.5}$, even though their study populations also consisted of Californian children.

PM_{10} was positively associated with ASD in the following studies:
Volk et al. [2] in 2013, Kalkbrenner et al. [55] in 2015, Kim et al. [56] in 2017, Chen et al. [50] in 2018.

In contrast, the following studies found no association between PM_{10} and ASD:
Gong et al. [3], Guxens et al. [53], Gong et al. [45], Ritz et al. [44], Yousefian et al. [57].

2.3.2.1.3 Ozone

Ozone **(**O_3) is a natural gas molecule made up of three oxygen atoms joined together. Ozone can be classified as Good and Bad Ozone. Good ozone is found naturally in the upper Stratosphere which forms an ozone layer around the Earth and protects from the sun's harmful ultraviolet radiation by absorbing them. On the other hand, Bad or ground-level ozone is found in the Troposphere where the lowest layer of Earth's atmosphere. It does not exist naturally but is mainly formed through chemical reactions between nitrogen oxides (NOx) and volatile organic compounds (VOC) when sunlight reacts with the pollutants emitted from the human activities (e.g, vehicles, power plants, factories, refineries, etc.).

Ground-level ozone is the major component of "smog". It is very harmful to humans and plants life because it contaminates the air and oxidizes biological tissues.

Someone can be exposed to unhealthy highest levels of Ozone on hot sunny days in urban areas while exercising or working outdoors in the middle of the day.

Exposure to ozone may damage the lungs and reduce lung function especially in developing children. Additionally, it may give damage to a fetus and may increases the risk of premature death, especially in people having heart and lung disease [58, 59, 60, 34, 37].

In, Becerra et al. [37] and Jung et al. [61] in 2013 studies, the relation between ASD and Ozon(O3) is found. However, Volk et al. [2], Kerin et al. [62], Kaufman et al. [63] found no associations between ASD and Ozon (O_3).

3. DATA

3.1 Data Collection

The data used in this thesis has been collected from the National Database repository for Autism Research (NDAR) data repository which is developed by The National Institutes of Health (NIH) [64].

Researchers requesting to access data contained in NDAR Collections, first of all, should submit data use certification DUC and this submission should be approved by Data Access Comite (DAC). This request procedure takes a little bit of time.

3.2 Datasets

The data documents which are in text format with the titles shown in Table 3.1 are downloaded from the NDAR data repository [64], by first logging in and clicking Data Dictionary [65] then searching and filtering for the "Exposure" category. Table 3.1 shows the information about the datasets in "Distance to Freeway and Major Road Data" Collection.

Table 3.1: Distance to Freeway and Major Road Data Collection

No	Datasets Title	Txt Files	Number of Subjects	Number Of Records
1	Traffic Related Air Pollution (TRAP) Estimates	trp_estimates01	1039	6206
2	Distance to Roadways	roaddistance02	1049	1049
3	Nitrogen Dioxide (NO_2) Exposure	no2_exposure01	1049	6284
4	Ozone(O_3) Exposures	o3_exposure01	1049	6284
5	Particulate Matter 10 (PM_{10}) Exposures	pm10_exposures01	1049	6284
6	Particulate Matter 2.5 ($PM_{2.5}$) Exposures	pm25_exposures01	1049	6284

The working data is a collection with the title “Distance to Freeway and Major Road” and whose investigator is Rob McConnell. The data in these collections belong to the CHARGE study of pregnant mothers. It contains the measure of some exposures estimated according to the distance to some roads. The subjects enrolled in the CHARGE study in the above data collections, shown below, includes also both pregnant mother and their child info. The phenotype of the child is associated with the mother's data. CHARGE subjects were preschool children between 24 and 60 months of age and were born between 1997 and 2006 in California when they joint with the organization.

The exposures measures were estimated for each trimester of pregnancy and the first year of life by using the mother’s address and the Environmental Protection Agency’s Air Quality System data [2].

The information contained in these datasets are:

- ***Traffic-related air pollution (TRAP) estimate***: dataset contains traffic-related air pollution (TRAP) estimate averages for 5 time periods. They are constructed based on mothers’ address locations. The concentrations of nitrogen oxides from freeways, non-freeways, and all roads located within 5 km of each child’s home are estimated by using the CALINE-4 line source dispersion model. Here, nitrogen oxides (NOx) concentrations can be viewed as an indicator of the TRAP mixture since they showed a perfect correlation with other traffic-related pollutants before [2].

- ***Distance to Roadways***: contains the distances (in meters) to nearest class-1, class-2, class-3, and class-4 roads to subjects birth residence(on birth address)
- ***Nitrogen Dioxide (NO$_2$) Exposure:*** contains NO$_2$ estimated averages for 5 time periods. Nitrogen oxides (NOx) concentrations can be viewed as an indicator of the TRAP mixture since they showed a perfect correlation with other traffic-related pollutants.

- ***Ozone (O$_3$) Exposures***: contains ozone estimated averages for 5 time periods

- ***Particulate Matter 10 (PM$_{10}$) Exposures***: contains PM$_{10}$ estimated averages for 5 time periods.

- ***Particulate Matter 2.5 (PM$_{2.5}$) Exposures***: contains PM$_{2.5}$ estimated averages for 5 times.

Where exposure during 6 time periods under study includes:

- ***Preg***: All pregnancy,
- ***Trim1***: First trimester,
- ***Trim2***: Second trimester,
- ***Trim3***: Third trimester.
- ***1stYr***: First year,
- ***2ndYr***: Second year.

and road type and their definitions are given as:

- Class-1: Interstate highway
- Class-2: the US and state highways
- Class-3: Secondary state or county highway
- Class-4: Local, neighborhood, rural road, city street

The exposure measurements for PM$_{2.5}$, PM$_{10}$, O$_{3,}$ and NO$_2$ were estimated from the US EPA's Air Quality System (AQS) data by using the regional air quality data [2].

3.3 Data fields of the Datasets

"*Traffic-Related Air Pollution (TRAP) Estimates*", "*Distance to Roadways*" "*Nitrogen Dioxide (NO$_2$) Exposures*", "*Ozone(O$_3$) Exposures*", "*Particulate Matter 10 (PM$_{10}$) Exposures*", "*Particulate Matter 2.5 (PM$_{2.5}$) Exposures*" dataset attributes were shown in Table 3.2, Table 2.3, Table 3.4, Table 2.5, Table 3.6, and Table 3.7 respectively.

Table 3.2: "Traffic-Related Air Pollution (TRAP) Estimates" Table Attributes

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	periodname	Time Period for Exposure(Preg,Trim1,Trim2,Trim3,1stYr,2ndYr)
5	roadtype1_nox	Avg. NO$_x$ in ppb estimate from Interstate Highways
6	roadtype2_nox	Avg. NO$_x$ in ppb estimate for US and State Highways
7	roadtype3_nox	Avg. NO$_x$ in ppb estimate for secondary state or county highways
8	roadtype4_nox	Avg. NO$_x$ in ppb estimate for local, neighborhood, rural road, or city street
9	roadtypeAll_nox	Avg. NO$_x$ in ppb estimate from all road types

Table 3.3: "Distance to Roadways" Table Attributes

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	fcc1_distance	Distance (in meters) to the nearest class-1 road to the birth residence
5	fcc2_distance	Distance (in meters) to the nearest class-2 road to the birth residence
6	fcc3_distance	Distance (in meters) to the nearest class-3 road to the birth residence
7	fcc4_distance	Distance (in meters) to the nearest class-4 road to the birth residence

Table 3.4: "Nitrogen Dioxide (NO$_2$) Exposures" Table Attributes

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	periodname	Time Period for Exposure(Preg,Trim1,Trim2,Trim3,1stYr,2ndYr)
5	no2_exposure	Avg. NO$_2$ in ppb

Table 3.5: "Ozone(O$_3$) Exposures" Table Attributes

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	periodname	Time Period for Exposure (Preg,Trim1,Trim2,Trim3,1stYr,2ndYr)
5	ozone_exposure	Avg. O$_3$ exposure from 8-hour daytime average, 10am - 6pm, ppb

Table 3.6: “Particulate Matter 10 (PM_{10}) Exposures” Table Attributes

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	periodname	Time Period for Exposure(Preg,Trim1,Trim2,Trim3,1stYr,2ndYr)
5	pm10	Avg. exposure to PM 10 in ppb

Table 3.7: “Particulate Matter 2.5 ($PM_{2.5}$) Exposures” Table Attributes

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	periodname	Time Period for Exposure(Preg,Trim1,Trim2,Trim3,1stYr,2ndYr)
5	pm25	Avg. PM 2.5 in ppb (part per billion)

All the tables in Table 3.1 were imported to new tables in Microsoft Access. Then, “*Nitrogen Dioxide (NO_2) Exposures*”, “*Ozone(O_3) Exposures*”, “*Particulate Matter 10 (PM_{10}) Exposures*” and “*Particulate Matter 2.5 ($PM_{2.5}$) Exposures*” tables were also transformed and transposed into new temporary tables such as tempNO2, tempO3, tempPM10 and temp25 respectively with a large number of rows but a small number of dimensions as in Table 3.8 and Table 3.9.

Table 3.8: Transformed “Nitrogen Dioxide (NO_2) Exposures” Table tempNO2

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	roadtype1_1stYr_nox	Avg. NOx in ppb estimate from class-1 road during 1st Year
5	roadtype1_2ndYr_nox	Avg. Nox in ppb estimate from class-1 during 2nd Year
6	roadtype1_Preg_nox	Avg. NOx in ppb estimate from class-1 during Pregnant
7	roadtype1_Trim1_nox	Avg. NOx in ppb estimate from class-1 during the First trimester
8	roadtype1_Trim2_nox	Avg. NOx in ppb estimate from class-1 during the Second trimester
9	roadtype1_Trim3_nox	Avg. NOx in ppb estimate from class-1 during the Third semester
10	roadtype2_1stYr_nox	Avg. NOx in ppb estimate for class-2 during 1st Year
11	roadtype2_2ndYr_nox	Avg. NOx in ppb estimate for class-2 during 2nd Year
12	roadtype2_Preg_nox	Avg. NOx in ppb estimate for class-2 during Pregnant
13	roadtype2_Trim1_nox	Avg. NOx in ppb estimate for class-2 during the First trimester
14	roadtype2_Trim2_nox	Avg. NOx in ppb estimate for class-2 during the Second trimester
15	roadtype2_Trim3_nox	Avg. NOx in ppb estimate for class-2 during the Third trimester
16	roadtype3_1stYr_nox	Avg. NOx in ppb estimated for class-3 during 1st Year

17	roadtype3_2ndYr_nox	Avg. NOx in ppb estimated for class-3 during 2nd Year
18	roadtype3_Preg_nox	Avg. NOx in ppb estimated for class-3 during Pregnant
19	roadtype3_Trim1_nox	Avg. NOx in ppb estimated for class-3 during the First trimester
20	roadtype3_Trim2_nox	Avg. NOx in ppb estimated for class-3 during the Second trimester
21	roadtype3_Trim3_nox	Avg. NOx in ppb estimated for class-3 during the Third trimester
22	roadtype4_1stYr_nox	Avg. NOx in ppb estimated for class-4 during 1st Year
23	roadtype4_2ndYr_nox	Avg. NOx in ppb estimated for class-4 during 2nd Year
24	roadtype4_Preg_nox	Avg. NOx in ppb estimated for class-4 during Pregnant
25	roadtype4_Trim1_nox	Avg. NOx in ppb estimated for class-4 during the First trimester
26	roadtype4_Trim2_nox	Avg. NOx in ppb estimated for class-4 during the Second trimester
27	roadtype4_Trim3_nox	Avg. NOx in ppb estimated for class-4 during the Third trimester
28	roadtypeAll_1stYr_nox	Avg. NOx in ppb estimated from all road types during 1st Year
29	roadtypeAll_2ndYr_nox	Avg. NOx in ppb estimated from all road types during 2nd Year
30	roadtypeAll_Preg_nox	Avg. NOx in ppb estimated from all road types during Pregnant
31	roadtypeAll_Trim1_nox	Avg. NOx in ppb estimated from all road types during the First trimester
32	roadtypeAll_Trim2_nox	Avg. NOx in ppb estimated from all road types during the Second trimester
33	roadtypeAll_Trim3_nox	Avg. NOx in ppb estimated from all road types during the Third trimester

Table 3.9: Transformed “Ozone(O_3) Exposures” Dataset tempO3

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	o3_2ndYr	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during 2nd Year
5	o3_Preg	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during Pregnant
6	o3_Trim1	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during the First trimester
7	o3_Trim2	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during the Second trimester
8	o3_Trim3	8 hour daytime(10am - 6pm) Avg. ozone exposure during the Third trimester

Then, a new “Exposure” table was created by joining tempNO2, tempO3, tempPM10, and temp25 tables with the “subject key” field and inserting their data into the “Exposure” table. The newly formed “Exposure” table shown in Table 3.10 consists of 1039 records or instances corresponding to a single subject. It contains a total of 61 separate data fields having 1 text, 2 categorical, 59 numerical fields. Details of the fields are given in Table 3.10.

Table 3.10: “Exposure” Dataset Attributes

No	Attribute Name	Description
1	Subject_key	The NDAR Global Unique Identifier (GUID) for research subject
2	gender	Gender
3	phenotype	Phenotype/diagnosis for the subject
4	fcc1_distance	Distance (in meters) to the nearest class-1 road to the birth residence
5	fcc2_distance	Distance (in meters) to the nearest class-2 road to the birth residence
6	fcc3_distance	Distance (in meters) to the nearest class-3 road to the birth residence
7	fcc4_distance	Distance (in meters) to the nearest class-4 road to the birth residence
8	roadtype1_1stYr_nox	Avg. NOx in ppb estimate from class-1 road during 1st Year
9	roadtype1_2ndYr_nox	Avg. Nox in ppb estimate from class-1 during 2nd Year
10	roadtype1_Preg_nox	Avg. NOx in ppb estimate from class-1 during Pregnant
11	roadtype1_Trim1_nox	Avg. NOx in ppb estimate from class-1 during the First trimester
12	roadtype1_Trim2_nox	Avg. NOx in ppb estimate from class-1 during the Second trimester
13	roadtype1_Trim3_nox	Avg. NOx in ppb estimate from class-1 during the Third semester
14	roadtype2_1stYr_nox	Avg. NOx in ppb estimate for class-2 during 1st Year
15	roadtype2_2ndYr_nox	Avg. NOx in ppb estimate for class-2 during 2nd Year
16	roadtype2_Preg_nox	Avg. NOx in ppb estimate for class-2 during Pregnant
17	roadtype2_Trim1_nox	Avg. NOx in ppb estimate for class-2 during the First trimester
18	roadtype2_Trim2_nox	Avg. NOx in ppb estimate for class-2 during the Second trimester
19	roadtype2_Trim3_nox	Avg. NOx in ppb estimate for class-2 during the Third trimester
20	roadtype3_1stYr_nox	Avg. NOx in ppb estimated for class-3 during 1st Year
21	roadtype3_2ndYr_nox	Avg. NOx in ppb estimated for class-3 during 2nd Year
22	roadtype3_Preg_nox	Avg. NOx in ppb estimated for class-3 during Pregnant
23	roadtype3_Trim1_nox	Avg. NOx in ppb estimated for class-3 during the First trimester
24	roadtype3_Trim2_nox	Avg. NOx in ppb estimated for class-3 during the Second trimester
25	roadtype3_Trim3_nox	Avg. NOx in ppb estimated for class-3 during the Third trimester
26	roadtype4_1stYr_nox	Avg. NOx in ppb estimated for class-4 during 1st Year
27	roadtype4_2ndYr_nox	Avg. NOx in ppb estimated for class-4 during 2nd Year
28	roadtype4_Preg_nox	Avg. NOx in ppb estimated for class-4 during Pregnant
29	roadtype4_Trim1_nox	Avg. NOx in ppb estimated for class-4 during the First trimester
30	roadtype4_Trim2_nox	Avg. NOx in ppb estimated for class-4 during the Second trimester
31	roadtype4_Trim3_nox	Avg. NOx in ppb estimated for class-4 during the Third trimester
32	roadtypeAll_1stYr_nox	Avg. NOx in ppb estimated from all road types during 1st Year
33	roadtypeAll_2ndYr_nox	Avg. NOx in ppb estimated from all road types during 2nd Year
34	roadtypeAll_Preg_nox	Avg. NOx in ppb estimated from all road types during Pregnant
35	roadtypeAll_Trim1_nox	Avg. NOx in ppb estimated from all road types during the First trimester
36	roadtypeAll_Trim2_nox	Avg. NOx in ppb estimated from all road types during the Second trimester
37	roadtypeAll_Trim3_nox	Avg. NOx in ppb estimated from all road types during the Third trimester
38	no2_1stYr	Avg. NO2 in ppb during 1st Year
39	no2_2ndYr	Avg. NO2 in ppb during 2nd Year
40	no2_Preg	Avg. NO2 in ppb during Pregnant
41	no2_Trim1	Avg. NO2 in ppb during the First trimester

42	no2_Trim2	Avg. NO2 in ppb during the Second trimester
43	no2_Trim3	Avg. NO2 in ppb during the Third trimester
44	o3_1stYr	Avg. ozone exposure in ppb during 1st Year
45	o3_2ndYr	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during 2nd Year
46	o3_Preg	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during Pregnant
47	o3_Trim1	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during the First trimester
48	o3_Trim2	8 hour daytime(10am - 6pm) Avg. ozone exposure in ppb during the Second trimester
49	o3_Trim3	8 hour daytime(10am - 6pm) Avg. ozone exposure during the Third trimester
50	pm10_1stYr	Average exposure to PM 10 in ppb during 1st Year
51	pm10_2ndYr	Average exposure to PM 10 in ppb during 2nd Year
52	pm10_Preg	Average exposure to PM 10 in ppb during Pregnant
53	pm10_Trim1	Average exposure to PM 10 in ppb during the First trimester
54	pm10_Trim2	Average exposure to PM 10 in ppb during the Second trimester
55	pm10_Trim3	Average exposure to PM 10 in ppb during the Third trimester
56	pm25_1stYr	Average PM 2.5 in ppb during 1st Year
57	pm25_2ndYr	Average PM 2.5 in ppb during 2nd Year
58	pm25_Preg	Average PM 2.5 in ppb during Pregnant
59	pm25_Trim1	Average PM 2.5 in ppb during the First trimester
60	pm25_Trim2	Average PM 2.5 in ppb during the Second trimester
61	pm25_Trim3	Average PM 2.5 in ppb during the Third trimester

The summary of the subjects in “Exposure” table was shown in Table 3.11

Table 3.11: The distribution of the subjects in “Exposure” Dataset

Phenotype	# of Subjects
AUTISM SPECTRUM AFFECTED	141
AUTISM SPECTRUM SEVERELY AFFECTED	338
NEUROLOGICAL CONTROL	136
NOT DEFINED	152
TYPICAL CONTROL	272
TOTAL	1039

There are some mechanisms and rules used for phenotype categorization when they are shared data in NDAR. Subjects Categorization is performed based upon the following order:

1. Fragile X
2. Controls

- Non-Spectrum Typical Control (e.g. typical, sibling, parent)
- Non-Spectrum Neurological Control

3. Autism Spectrum

- Severely Affected
- Mildly Affected
- Affected

Fragile X is defined according to provided genetic test results for the Fragile X mutation of the FMR1 gene.

Typical controls are typically developing individuals. The Neurological disorders sub-phenotype control group includes subjects with a learning disability, Attention Deficit Hyperactivity Disorder, developmental disability, intellectual disability/MR, or other neurological disorders, excluding Fragile X and subjects with positive genetic test result for Non-Spectrum Neurological conditions.

NDAR categorizes Typical and Neurological Disorder control subjects based on results from the ADI-R, ADOS, IQ, and Vineland Survey assessments. NDAR also categorizes the AUTISM SPECTRUM AFFECTED, AUTISM SPECTRUM SEVERELY AFFECTED phenotypes according to cut-offs, for each Assessment (ADI-R, ADOS, IQ, and Vineland Survey). Note that a minimum of three assessments -including ADI-R and ADOS – plus one other measure (Vineland or an IQ) is needed for the categorization of an autism spectrum phenotype. 'Not defined" means that not enough data provided to define phenotype. In the absence of a diagnosis, NDAR categorizes control subjects based on the results from the ADI-R, ADOS, IQ, and Vineland Survey assessments.

4. METHODS

4.1 Data Preprocessing

Data preprocessing is an important phase of the data mining process which aims to improve data quality and efficiency of the data mining process. Since the real world row data is not perfect and dirty, it modifies the data to make it more suitable and transforms raw data into an understandable format for data analysis [66, 13].

Most of the analyzed raw data have generally common data quality problems

- ***Completeness***: Not all attributes in the data table have correct values for missing value and some records might have missing values because of some technical reasons. For example, a survey might be filled out by skipping age information.
- ***Accuracy***: It refers to the deviation of the data value from the true or expected value. For numerical attributes, out of range values or the reduction in the correctness of the data can be caused by noise or wrong measurements.
- ***Inconsistent***: Inconsistent values in data occur when entering wrong codes or information instead of what should be. Let's say, the user entered birthday to be May 07, 1993, and the age attribute displays 50.

The major steps involved in data preprocessing are data cleaning and data transformation. They are useful for the databases that need to preprocessing. Data preprocessing can be responsible for 10–60% of time and effort in the whole data mining process [13].

4.1.1 Data Cleaning

Data cleaning is a process that improves data quality. It involves filling in missing values, correction of simple errors, correcting inconsistencies, removing noise, duplicate records, and outliers [67, 66]. It is a necessary time-consuming procedure and needs serious effort for successful data mining [68].

4.1.1.1 Handling Missing Values

Missing data may stem from many different causes. While analyzing data, it is quite often encountered that no data value is stored in certain records for some of the attributes quite often occur in datasets. For example, a sensor may be defective and may not send healthy data, or some participants in a survey may refuse to answer or skips some questions, or mistakes are made in data entry when data collection is done improperly [67].

Missing data may become a serious problem as the data analysis becomes more complex. Data analysts have to be deal with missing values. If the missing values are not handled properly by the analyst, then inaccurate inferences about the data or false conclusions can be made at the end.

An important question is "How can we deal with the attributes having missing values ?" Several strategies in handling a dataset with missing value can be followed, especially replacing the missing value with a value according to various criteria [13].
Some common methods are as follows:

- ***Deleting records(rows) with unknowns***: This is usually done when the record contains several fields with missing values.

- ***Dropping attributes with unknowns***: This is usually done when the particular attribute(variable) contains more missing values that the rest of the variables in the dataset.

- **Replace missing value with a constant**: Generally, an analyst decides which constant to be replaced with missing values. A common choice is to replace all missing values by the same constant such as *"unknown"*, -∞ , or NA (not available). In R, it is supported by many functions such as sum, prod, quantile, and sd through the na.rm option and a missing value is represented by NA (not available).

- **Replace the missing value with Measures of Central Tendency:** The most commonly used measures of central tendencies are the mean, median, and mode. Missing values for a given attribute are replaced by mean, median, or mode For normal data distributions, especially for symmetric data, mean can be used. But, for skewed data distribution, the median can be used**.**
- **Replace the missing values with the most probable value**: This may be determined with imputed values based on the other characteristics of the record. The question "What would be the most likely value for this missing value, given all the other attributes for a particular record?" should be answered. [66, 67].

4.1.1.2 Handling Noisy Data

When we say ***noisy***, it refers to the change of a value, the addition of meaningless data or out of range values like a person filling out the numeric value -679 in the salary field or some negative four-digit random number in the age field. Noise is one of the random problems which is involved in measurement error. The process of removing noise from a dataset is termed as data smoothing. The following techniques can be used for noise reduction and data smoothing:

- **Binning:** It is a technique where the data is sorted and then partitioned into equal frequency bins. Then the noisy data may either be replaced with the bin mean bin median, or the bin boundary.
- **Regression:** Regression is used to find a mathematical equation to fit the data which helps to smooth out the noise [66].

4.1.1.3 Outliers Detection and Treatment

In Data Science, an *outlier* is simply an extreme value or data objects that have different characteristics from most of the other data objects in the dataset [69]. Causes of outlier may be due to measurement errors, incorrect data entry, or incorrect selection of a sample [67]. Figure 4.1 below provides a visual understanding of *Outliers.* According to Figure 4.1, the objects in region R can be identified as an outlier. Because, the other group of objects in the dataset is close to each other, falls into the same cluster, and follow the same distribution [66].

If there are outliers in the dataset, they can drastically change the results of the data analysis and give unreliable results. For example, they can increase the variance, decrease normality, and reduces the reliability of the tests [13, 70].

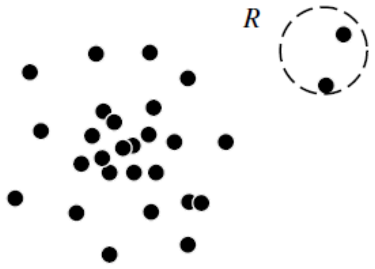


Figure 4.1: The objects in region *R* are outliers.

4.1.1.3.1 Outlier Detection

Outlier detection is the most important process of finding *anomalies.* It is used in many applications such as fraud detection, or image processing. There are two outlier detection methods which are called *univariate and multivariate.* While the *univariate* method detects outliers on one variable, on the other hand, the *multivariate* method detects unusual combinations on all the variables.

In our analyses, we will be concerned with *The Box Plot Rule* [71] for outlier detection in univariate numerical data. It is a graphical tool used to construct a boxplot for more or less unimodal and symmetrically distributed data and to display a dataset based on the five-number summary, such as the minimum, first quartile, median, third quartile, and maximum. In this method, the interquartile range (IQR) is used to find outliers and to filter out very large or small numbers. To create a boxplot as in Figure 4.2, the rules of the method are as: Firstly order the data from smallest to largest. Secondly, find the median, the first quartile(Q1), the third quartile(Q3), min, and max of the data. Then calculate the IQR which is the difference between the first and third quartile values. (Q3 - Q1). Next, calculate the lower fence, 1.5 X IQR. Q3 - [(IQR) x 1.5], and calculate the upper fence, 1.5 X IQR. Q3 + [(IQR) x 1.5]. Then, draw and label the axes of the graph and a box from Q_1 to Q_3 with a vertical line through the median. Finally, draw a whisker from Q_1 to the min and from Q_3 to the max.

As a result, an outlier is any value that lies more than the upper fence or below the lower fence. Outliers lie outside the fences, that is, if a data point is below $Q_1 – 1.5×IQR$ or above $Q_3 + 1.5×IQR$ [72]. If we consider "extreme values", they are the values lies between Min and $Q_1 – 3×IQR$ and Max and $Q_3 + 3×IQR$. The outliers are marked with asterisks(*) and extreme values are X [73].

Finally, it is assumed that values are normally clustered around some central value. The IQR demonstrates how the middle values spread out and how too far some of the other values from the central value are. These "too far" points are called "outliers" because they "lie outside" the range in which we expect them [73].

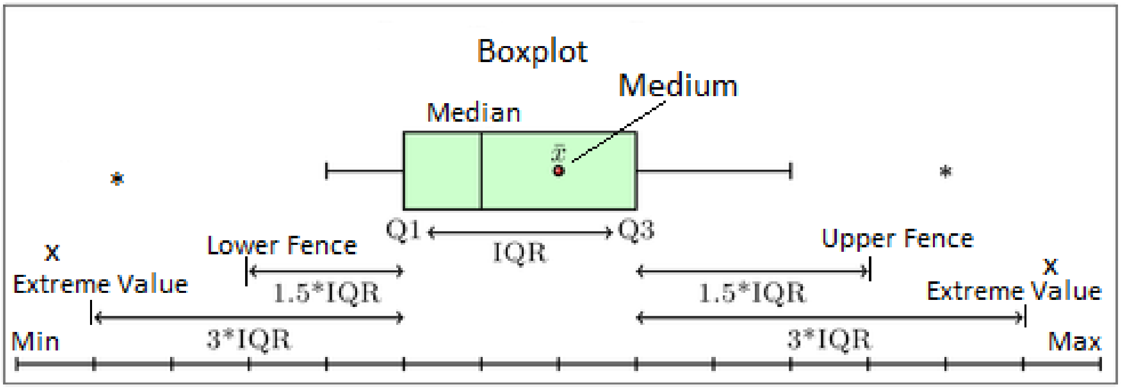


Figure 4.2: Boxplot

Additionally, histograms can also be used as a graphical method to identify outliers for numeric variables.

4.1.1.3.2 Removing Outliers

There are lots of strategies for dealing with deal with outliers. They are similar to the methods of missing values. In this research, the strategy of how to remove and when to replace outliers depends on the distribution of the data. Most of the data in applications are not symmetric, that is, they are either positively skewed or negatively skewed shown in Figure 4.3. Therefore, if we have approximately normal (symmetrical) distributions for continuous data, where all observations are nicely clustered around the mean which is a good option. However, for skewed distributions shown in Figure 4.3, the median is the best choice in dealing with missing values [66].

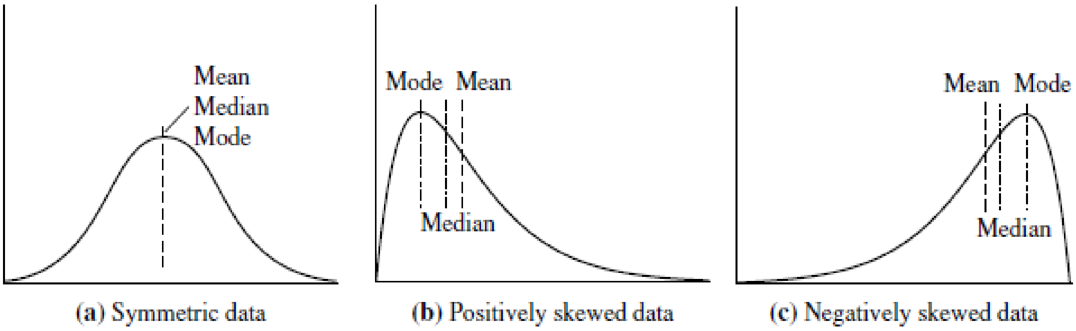


Figure 4.3: Mean, median, and mode of symmetric versus positively and negatively skewed data.

4.2 Logistic Regression

4.2.1 Introduction

Regression methods are an essential component of the data analysis to examine the relationship between the response variable and one or more independent variables of interest. These methods aim to find the best fitting for describing the relationship between dependent and independent variables. Logistic regression or logit model is the most frequently technique used when the dependent variable taking two or more values. The main types of logistic regression are:

- **Binary or Binomial Logistic Regression**: dealing with the cases in which the response variable can take only two 2 possible values (e.g. “0 / 1”, “dead/alive”, “win/loss”, “pass/fail”, etc.)
- **Multinomial logistic regression:** dealing with cases when the response can have 3 or more possible values that are not ordered (e.g., “TypeA/TypeB/TypeC”).
- **Ordinal logistic regression**: dealing with response variables that are ordered.

Binary Logistic regression is generally preferred for the analysis of binary responses in biological or social sciences and medical or epidemiologic researches, for example, to estimate the presence or absence of a particular disease, the effect of a treatment on a patient, or whether a firm will go bankrupt or not in a year [74].

Logistic regression can be defined as simple logistic regression with one independent variable and multiple logistic regression with more than one independent variable which may be either continuous or categorical.

Linear and Logistic regressions are similar to each other but also there are two main differences: Firstly, the response variable in logistic regression is binary while in linear regression the response variable is continuous. Secondly, while the outcome is binomially distributed, being in either group in logistic regression, it is normally distributed in linear regression [75]. We can give an illustrative example to show the similarities and differences between logistic and linear regression by using the data in Table 1.1 [75] listing the age and coronary heart disease (CHD) status for 100 subjects [75]. CHD indicates the response variable which is 1 if CHD is present in the individual, otherwise 0. The scatter plot of these data is given in Figure 4.4 shows no functional relationship between the observed values of AGE and response variable CHD and two parallel lines corresponding to the values of a binary response variable, does not provide enough information about the relationship between CHD and AGE and are also difficult to be described with linear regression.

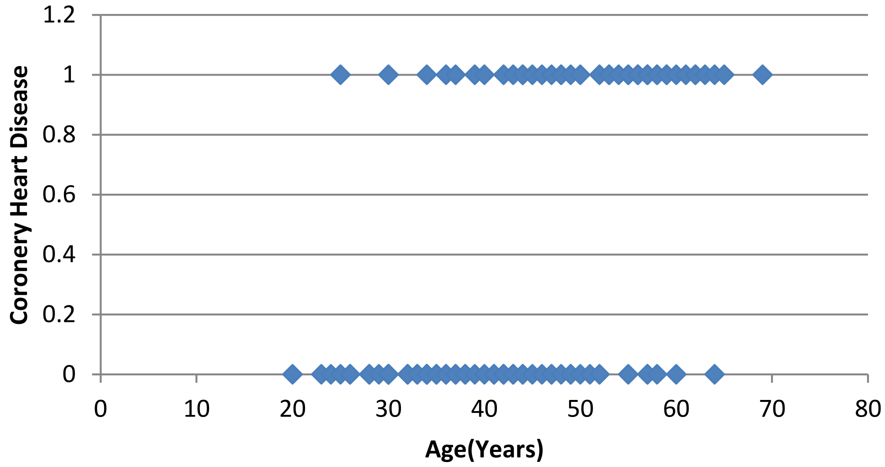


Figure 4.4: Scatterplot of coronary heart disease (CHD) status by age for 100 subjects

The strategy, by grouping age into the categories (AGEGRP) and for each age group by computing the number of occurrences and the mean(proportion) of the outcome variable shown in Table 4.1, can be used to overcome this problem.

If a graph of the mean of individuals in each group versus the midpoint of each age interval is plotted, it can be shown in Figure 4.5 that the probability of *CHD* increases with the *AGE* and thus, the relationship between *CHD* and *AGE* is nonlinear does not lie outside the range from 0 to 1.

Table 4.1: Frequency Table of Age Group by with CHD or without in each group

Age Group	# of individual in each group	with CHD	without CHD	Mean
20–29	10	1	9	0.10
30–34	15	2	13	0.13
35–39	12	3	9	0.25
40–44	15	5	10	0.33
45–49	13	6	7	0.46
50–54	8	5	3	0.63
55–59	17	13	4	0.76
60–69	10	8	2	0.80
Total	100	43	57	0.43

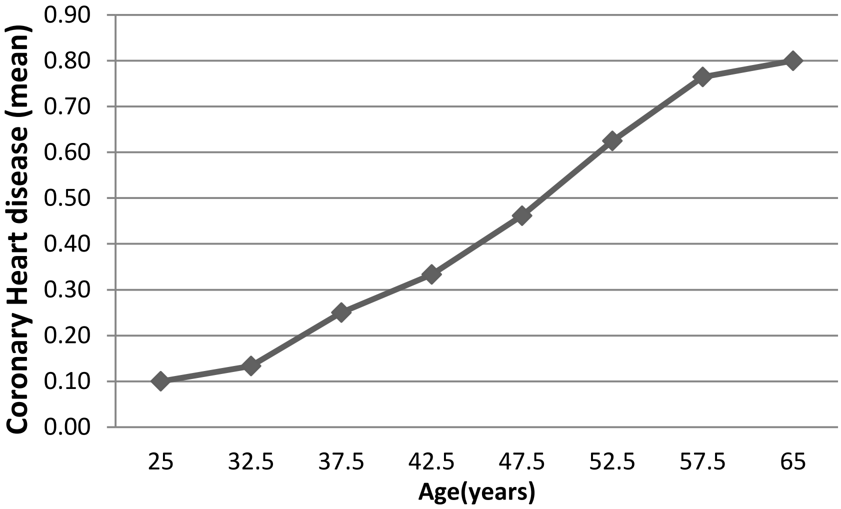


Figure 4. 5: Plot of the percentage of subjects with *CHD* in each AGE group.

The shape for the relationship displayed in Figure 4.5 is said to be an S-shaped curve. It resembles the plot of the logistic function shown in Figure 4.3, which is the most important mathematical function having the following exponential formula

$$\pi(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)} \tag{4.1}$$

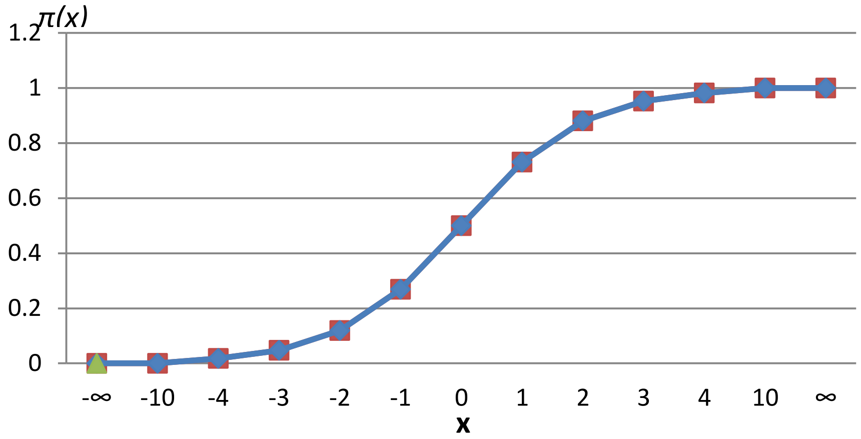


Figure 4.6: A graph of Logistic Function

As we see from the graph of the logistic function in Figure 4.6, the probability values of this function π(x) change very little at the low and or increases gradually and is restricted to range between 0 and 1 as x varies in the interval (-∞,+∞).

To describe the relationship between response and independent variables in logistic regression, the same techniques in linear regression can be used as well. So, in linear regression, the relationship between response variable *Y* and independent variable *X* is mathematically linear and can be simply described by the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{4.2}$$

where:

Y_i is the response variable value, β_0 and β_1 are unknown coefficients, X_i is the independent value and ε_i is the random error variable related to the *i*th subject. Indeed, another way of expressing Eq. 4.2 in terms of conditional expectation is as follows:

$$E(Y_i) = \beta_0 + \beta_1 X_i \tag{4.3}$$

where $E(Y_i)$ *is* the expected value of Y_i for the given each value of X_i .

E(Y_i) can get any values as x ranges between −∞ and +∞ in linear regression. But, the expectation of *E*(Y_i) has a special meaning when the response variable *Y* is binary taking on the value of either 0 or 1 with probabilities π and 1 − π respectively. Let's assume that Y_i is a Bernoulli random variable with the probability distribution as follows:

Y_i	Probability
1	$\pi_i = P(Y_i = 1)$
0	$1 - \pi_i = P(Y_i = 0)$

The expected value of the response variable is

$$E(Y_i) = 1\; x\; \pi_i + \; 0\; x\; 1 - \pi_i = \pi_i \tag{4.4}$$

This implies that

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i \tag{4.5}$$

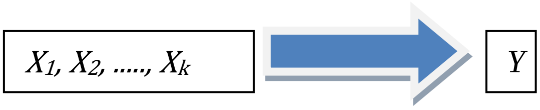
Thus, it can be concluded that the expected response was given by the response function $E(Y_i) = \beta_0 + \beta_1 X_i$ is just the probability that the response variable Y_i takes on the value 1. Since there is a constraint on the expected response $E(Y_i)$ such as $0 \leq E(Y_i) \leq 1$, the linear model in Eq. 4.2 will NOT work for the relationship CHD and AGE. Then, one of the ways of modeling the data when the response variable is binary is to use the logistic function. So, the simple logistic regression model can be stated for the i^{th} observation in the following fashion:

$$E(Y_i) = \; \pi_i = \frac{\exp(\beta_0 + \cdots . \beta_i X_i)}{1 + \exp(\beta_0 + \cdots \beta_i X_i)} \tag{4.6}$$

Alternatively $E(Y_i)$ is viewed as a conditional mean, given the value of X_i and take values between 0 and 1 (i.e., $0 \leq E(Y_i) \leq 1$).

Logistic Regression deals with the case where the dependent variable is binary, and the conditional distribution is binomial [76, 75]. The primary reason the logistic model to be popular is that it ranges between 0 and 1. As a result, the model can be designed to describe a probability, which is always some number between 0 and 1 and such a probability can also give the risk of an individual getting a disease in epidemiologic researches [77].

4.2.2 The Multiple Logistic Regression Model



Relating k number of independent variables denoted as X_1, X_2, ... X_k to dependent variable Y can be defined as a multivariable problem. The simple linear regression model (4.3) can easily be extended to multivariable linear with more than one predictor variable such that

$$E(Y \mid X) = \beta X = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad (4.7)$$

But, whenever we wish to relate a set of k independent variables denoted as X_1, X_2, ... X_k to a binary dependent variable Y, The simple logistic regression model in Eq. (4.6) can be extended by adding more than one independent variable to get multiple logistic regression. So, the preferred model for the analysis of binary responses is the multiple logistic regression model which can be stated in terms of the probability that $Y = 1$ for given a set of X, the values of the independent variables [74]:

$$Pr(Y = 1|\, X) = \frac{1}{1 + exp(-\beta X_i)} \qquad (4.8)$$

where βX stands for $\beta_1 X_1 + \cdots + \beta_k X_k$, $X_1 + \cdots + X_k$ are the independent variables and $\beta_0 + \beta_1 + \cdots + \beta_k$ are the unknown coefficients of k independent variables. By using $\pi = E(Y/X) = Pr(Y = 1/X)$ to represent conditional probability of Y given X. Therefore, Eq. 4.8 can be written as a simple notation as

$$\pi = \frac{1}{1 + exp(-\beta X_i)} \qquad (4.9)$$

4.2.3 Model Assumptions

If the logistic regression model is simply stated in the usual form:

$$Y = Pr(Y = 1|\, X\,) = E(Y \;=\; 1|X\,) = \beta X + \varepsilon \qquad (4.10)$$

Firstly, the dependent variables Y should be binary and have a binomial distribution. Secondly, independent variables should not be highly correlated with each other, that is, the model should have little or no multicollinearity. Thirdly, the error terms ε

depending on Y should also be binomially distributed. Finally, the independent variables are linearly related to the log odds. The logistic model assumptions are easily understood when this log transformation is made [74, 76].

4.2.4 Logit Transformation

A logistic model can be written in an alternative form which is also known as the logit model. Logit models can be achieved by the transformation of probabilities (π) such as:

$$logit(\pi) = ln\left[\frac{\pi}{1-\pi}\right] = X\beta \tag{4.11}$$

where $\pi = [1 \ + exp(-\beta X)]^{-1}$.

Then, the ratio $\frac{\pi}{1-\pi}$ can be written as

$$\frac{\pi}{1-\pi} = \frac{[1 \ +exp(-\beta X)]^{-1}}{1-[1 \ +exp(-\beta X)]^{-1}} = exp(\beta X) \tag{4.12}$$

After taking the natural log of expression (4.5), the expression (4.4) can be written as

$$ln\left[\frac{\pi}{1-\pi}\right] = ln[exp(\beta X)] = \beta X = (\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p) \tag{4.13}$$

Thus, in the logistic model, the logit transformation which is called the link function of the dependent variable yields a linear function of independent variables ($X\beta$).

Thus, the logit transformation of the logistic model in Eq. (4.6) produces linearity with the independent variables ($X\beta$). Then it becomes a special case of a General Linear Model (GLM). The logit of π, symbolized by "$logit(\pi)$", is also called a link function. Logistic regression models are often called logit models [78].

$$logit(\pi) = (\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p) \tag{4.14}$$

The logistic model is generally described in terms of its logit form, $logit(\pi)$ rather than in its original form as in Eq. 4.6.

In particular, the quantity $\log(\frac{\pi}{1-\pi})$ describes the *log-odds* [77]. “In its simplest form, Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occurring is π, then the probability of an event not occurring is $(1-\pi)$” [77].

Then, the formula for *odds* is therefore given by

$$Odds\{event\} = \frac{\pi}{1-\pi} \tag{4.15}$$

Finally, the logit-model can be written in terms of log-odds

$$logit(\pi) = log\left[\frac{\pi}{1-\pi}\right] = log(odds) \tag{4.16}$$

4.2.5 Model Fitting

The parameters in the logistic regression model are estimated using the maximum likelihood estimation (MLE) method [74], which is different from the linear regression estimation way [74]. To apply the maximum likelihood concept, we first need to construct the likelihood function that represents the joint probability or likelihood of the observed data or a sample [76, 77]. The parameters estimated by the maximum likelihood methods are the values that maximize this function and are those that agree most closely with the observed data [75].

The likelihood function can be constructed for given independent observations $x_1, x_2, \dots., x_n$ with the distribution function *f(x; θ)* where θ is a parameter of the distribution as

$$L(x_1, x_2, \dots., x_n; \theta) = f(x;\ \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) \tag{4.17}$$

Now, the above quantity $L(x_1, x_2, \dots., x_n; \theta)$ represents the likelihood of the sample which is the following joint probability of obtaining the sample values $x_1, x_2, \dots., x_n$

$$P(X_1 = x_1,\ X_2 = x_2\ , \dots \dots, X_n = x_n |\ \theta) \tag{4.18}$$

where $X_1, X_2, \dots., X_n$ denote the independent variables.

Now, for the binary logistic model, let's begin to develop the joint probability function of a sampling in which each individual has the same probability and denoting the response and the probability of response of the *i*th subject by Y_i and and π_i , respectively, the model states that

$$\pi_i = Pr(Y_i = 1| X_i) = \frac{1}{1 + exp(-\beta X_i)} \tag{4.19}$$

where: $Pr(Y_i = 1) = \pi_i$

Since each random variable Y_i has binomial distribution, the probability distribution function of an observed response Y_i for the given predictors X_i is as follows:

$$\pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \tag{4.20}$$

For the observed data $Y_1, Y_2, \dots., Y_n$, the joint probability function of the responses is the product of these probabilities for *i = 1,..., n.*

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \tag{4.21}$$

However, mathematically, it is easier to work with the natural log. So, taking the natural log of Eq. 4.21 yields the *log-likelihood* function :

$$l(\beta) = ln(L(\beta)) = \ln \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \tag{4.22}$$

$$= \sum_{i=1}^{n} [Y_i \, ln \, \pi_i + (1 - Y_i) \, ln(1-\pi_i)] \tag{4.23}$$

$$= \sum_{i=1}^{n} [Y_i \, ln(\frac{\pi_i}{1-\pi_i})] + \sum_{i=1}^{n} ln(1-\pi_i) \tag{4.24}$$

It follows from Eq. 4.19:

$$1 - \pi_i = [1 \; + exp(\beta X_i)]^{-1} \tag{4.25}$$

The *log-likelihood* function in (4.24) is rewritten by using the definition Eq. 4.19 of π_i and Eq. 4.13 above to allow them to be recognized as a function of the unknown parameters β as follows :

$$l(\beta) = \sum_{i=1}^{n} Y_i(\beta X_i) - \sum_{i=1}^{n} ln \, [1 + exp(\beta X_i)] \tag{4.26}$$

where $l(\beta)$ can be viewed as the *log-likelihood* function of parameters to be estimated, $\beta = (\beta_0, \beta_1, .. \beta_{k-1})$, for the given n observations $X_i = (X_{i1}, X_{i2}, \dots, X_{i(k-1)})$.

Once the likelihood function has been determined for a given set of study data, the method of maximum likelihood chooses that estimator of the set of unknown parameters β which maximizes the likelihood function $l(\beta)$ [77].

To find the unknown parameters β that maximizes likelihood function $l(\beta)$, *log-likelihood* function $l(\beta)$ must be differentiated for the coefficients $\beta_0, \beta_1, \ldots. \beta_{k-1}$ and the results of the first derivative equations must be set to zero as follows:

$$\frac{\partial l(\beta)}{\partial \beta_p} = \sum_{i=1}^{n} Y_i X_{ip} - \pi_i X_{ip} = 0 \qquad (4.27)$$

$$\sum_{i=1}^{n} [Y_i - \pi_i] = 0 \qquad (4.28)$$

And

$$\sum_{i=1}^{n} X_{ip} [Y_i - \pi_i] = 0 \qquad (4.29)$$

For each p between 1and (k-1)

Thus, the above equations must be solved for each β_p iteratively. After estimation of the parameters $\hat{\beta} = (b_0, b_1, \ldots. b_{k-1})$, which is the solution to these equations, can be substituted into the response function in Eq. 4.19 to obtain the below-fitted response function [76].

$$\hat{\pi}_i = \left[1 + exp\,(-\hat{\beta}X_i\right]^{-1} = [1 + e^{-(b_0 + b_1 X_{i1} + .., + b_{k-1} X_{i(k-1)})}]^{-1} \qquad (4.30)$$

where $\hat{\pi}_i$ is used to denote the fitted value for the *i*th case. The general fitted logistic response function is as follows:

$$\hat{\pi} = \frac{1}{1+e^{\hat{\beta}X}} \qquad (4.31)$$

where $\hat{\beta}X_i$ stands for $b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots \ldots \ldots + b_{k-1} X_{i(k-1)}$

4.2.6 Interpretation of the Model Coefficients

We need to provide some interpretation for coefficients *β₀, β₁, β₂, ... βₖ₋₁* in terms of *odds* and *log(odds)* from logit function and to answer the question "What do the coefficients in model tell us about the study?". Hence, the interpretation involves determining the functional relationship between the dependent variable and independent variables.

The proper interpretation of coefficients depends on the difference between two logit values. For example, in the simple logistic regression model, $logit(\pi)=\beta_0+\beta_1X$, when the independent variable changes only one unit, from $x=0$ to $x=1$, the difference between two logits as in Eq. 4.32 is equal to β_1, the coefficient of the single independent variable.

$$logit(\pi_1) - logit(\pi_0) = (\beta_0 + \beta_1 x\ 1) - (\beta_0 + \beta_1 x\ 0) = \beta_1 \qquad (4.32)$$

where $\pi_1 = \Pr(Y = 1 | X = 1)$ and $\pi_0 = \Pr(Y = 1 | X = 0)$.

Now, we need to introduce another important parameter like the odds ratio as a measure of association. Firstly, odds can be defined as the ratio of the probability of success of an event to the non-success (y = 0). This relationship can be shown as

$$Odds = \frac{\pi}{1-\pi} \qquad (4.33)$$

The relationship between the probability of success, π , and odds is shown in Table 4.2.

Table 4.2 Probability(**π**) -Odds Relation

π	**Odds**
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Then, an *odds ratio(OR)* is defined as the ratio of the odds of two different events. For example, for two *events A* and *B*, the formula of the odds ratio(OR) is

$$odds\ ratio(OR) = \frac{odds_A}{odds_B} = \frac{\frac{\pi_A}{(1-\pi_A)}}{\frac{\pi_B}{(1-\pi_B)}} \qquad (4.34)$$

An OR is generally used to measure the relationship between exposure and disease in epidemiologic studies. It allows us to see which an event more likely occur under certain situations. If OR > 1, then event A can be more likely, on the other hand
If OR < 1, then event B can be more likely.

By using the difference between two logit values in the simple logistic model with a single variable, *log (odds)=$\beta_0+\beta_1X$,* the *log odds* ratio can be given by

$$log(odds_1) - log(odds_0) = \log\left(\frac{odds_A}{odds_B}\right) = log(OR) \tag{4.35}$$

$$(\beta_0 + \beta_1 x\ 1) - (\beta_0 + \beta_1 x\ 0) = (\beta_0 + \beta_1 - \beta_0) = \beta_1$$

Therefore,

$$log(OR) = \beta_1 \tag{4.36}$$

After exponentiating both sides, the relationship between the odds ratio and the regression coefficient β_1 is found to be

$$OR = e^{\beta_1} \tag{4.37}$$

This case provides the conceptual foundation for all the other situations. Additionally, if we test what happens to the logit change in a multivariate model when one of the variables, *X*, varies while keeping others fixed. For example, if our model has three independent variables such as X_1, X_2, and X_3. It can be asked what happens to the logit while X_1 and X_2 stay constant and X_3 changes from 0 to 1.

Our model is $logit(\pi) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 = C + \beta_3X_3$
where $C = \beta_0 + \beta_1X_1 + \beta_2X_2$

We write two *logit* equation for two cases

Case-1: $logit(\pi_1) = log(odds_1) \ = C + \beta_3 x\ 1 = C + \beta_3$

Case-2: $logit(\pi_0) = \ log(odds_0) = C + \beta_3 x\ 0 = C$

When we take the logit difference of two cases,

$$\Delta logit = logit(odds_1) - logit(odds_0) = C + \beta_3 - C = \beta_3,$$

we obtain

$$\Delta logit = \beta_3.$$

Thus, it can be generalized for a multivariable model with *k* variable that the coefficient β_i of any focused variable X_i, represents the change in the log odds for per unit change in a single variable X_i when all other variables are held constant. The difference between the two fitted logit values can be expressed as follows

$$log(odds_1) - \ log(odds_2) = \log(\frac{odds_1}{odds_2}) = \beta_i \tag{4.38}$$

$$log(OR) = \beta_i \tag{4.39}$$

Where $\frac{odds_1}{odds_2} = \frac{odds\{Y=1 \mid X_1,X_2,\ldots,X_i=1,\ldots\ldots\ldots X_k\}}{odds\{Y=1 \mid X_1,X_2,\ldots,X_i=0,\ldots\ldots\ldots X_k\}}$

Thus, the coefficient, β_i , is the logit difference or the difference between two log-odds when the vale of independent variable X_i = 1 or 0. In practice, it is hard to explain the log-odds. But, to provide the *odds ratio* as a measure of association is more meaningful for interpretation. So, taking antilogs of each side of Eq. 4.39, we see that the *OR* equals to the exponential function of the regression coefficient β_i of i^{th} independent variable [75, 74].

$$OR = \frac{odds_1}{odds_2} = e^{\beta_i} \tag{4.40}$$

If the odds measures exposure-disease relationship, it determines the strength of association as weak (OR=1 - 1.5), moderate (OR=1.51 - 2.5), and strong (OR>2.5) [79].

On the other hand, if we increase X_i from *m* to (*m* + d) while X_1 and X_2 stay constant, then, two *logit* equation for the example cases are:

Case-1: $logit(\pi_{\text{m+d}}) = log(odds_{\text{m+1}}) \ = C + \beta_3 * (\text{ m + d}) = C + \beta_3 * \text{m} + \beta_3 * \text{d}$

Case-2: $logit(\pi_m) \quad = log(odds_m) \quad\ = C + \beta_3 * \ m$

Then, the difference between two *logits* is

$$\Delta logit = log(odds_{m+d}) - log(odds_m) = C + \beta_3 * d - C = \beta_3 * d$$

$$log(\frac{odds_{\text{m+d}}}{odds_m}) = \beta_3 * d \tag{4.41}$$

Therefore, by exponentiating both sides of Eq. 4.41, the OR, equals to

$$OR = \frac{odds\ when\ x_i=m+d}{odds\ when\ x_i=m} = e^{\beta_i * d} = (e^{\beta_i})^d \tag{4.42}$$

4.2.7 Testing for the Significance of Coefficients

After estimating the coefficients, obtaining the ML estimates, our first concern is the assessment of the significance of the variables in the model. The *Wald test* and the *likelihood ratio test (LRT)* are commonly used to test the significance of regression parameters in a standard logistic regression [77]. Inference based on the Wald statistic is simplest, but the likelihood-ratio inference is more trustworthy [78].

4.2.7.1 Likelihood Ratio Test: Test Whether Several $\beta_k = 0$

The likelihood ratio (LR) significance test is analogous to F-test for linear models. It is based on a comparison of two "nested models", one model is considered a special case or subset of another model, with maximum likelihood estimation. It can be used to assess the contribution of individual predictors to a given model. It may be helpful to compare models with the LRT to see if additional terms are significant or not. The LRT is also used in determining whether a subset of the X variables in a multiple logistic regression model can be dropped or not, that is, testing whether the associated regression coefficients β_k equal zero. To illustrate how the significance of regression parameters in multiple logistic regression are tested, we consider the following M_F to M_R models to compare. We refer to M_F as the full model and to M_R as the reduced model which is obtained by setting certain parameters in the full model equal to zero [76, 80].

We begin with response functions for the full and reduced logistic model:

$$Model\ M_F:\ logit\ (\pi_F) = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_q X_q + \cdots + \beta_p X_p\ (Full\ Model)$$

$$Model\ M_R:\ logit\ (\pi_R) = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_q X_q \qquad (Reduced\ Model)$$

where β_0, β_1, up through β_p or β_q denotes the parameters to be estimated in the models and $p\ and\ q$ are the numbers of parameters in models $M_F\ and\ M_R$ respectively. Then, we first find the maximum likelihood estimates $b_F = (b_0, b_1, \ldots, b_p)$ for the full model containing *p parameters* and evaluate the likelihood function $L(\beta_F)$ in Eq. 4.13 when $\beta_F = b_F$. We denote this value of the likelihood function for the full model by L_F. Next, we obtain the maximum likelihood estimates $b_R = (b_0, b_1, \dots, b_q)$ for the reduced model and evaluate the likelihood function, $L(\beta_R)$ for the reduced model containing *q parameters* when $\beta_R = b_R$. We denote this value of the likelihood function for the reduced model by L_R.

The hypothesis we wish to test is:

$$H_0: \beta_{q+1} = \beta_{q+2} = \dots\dots\dots\dots\dots . = \beta_p = 0 \qquad (4.43a)$$

$$H_1: \beta_{q+1} \neq 0, \beta_{q+2} \neq 0, \dots\dots\dots ., \beta_p \neq 0 \qquad (4.43b)$$

The difference between log-likelihood for two models is called a likelihood ratio (LR), its test static is denoted by χ^2 shown as:

$$\chi^2 = -2LL = -2ln(likelihood\ Ratio) = -2\ ln\left[\frac{L_R}{L_F}\right] = -2[ln\,L_R - lnL_F] \qquad (4.44)$$

Since a model having more parameters better fits the data, the relation between the maximized likelihood values can be written as $L_R \leq L_F$ and therefore

$$-2ln\,L_F \leq -2ln\,L_R.$$

The statistic $-2ln\,L_R$ is called the *log-likelihood* for M_R and similarly $-2ln\,L_F$ for M_F. If $L_F \gg L_R$, then ratio $\frac{L_R}{L_F}$ approaches 0 and $-2ln\,[\frac{L_R}{L_F}]$ approaches +∞. If $L_R = L_F$, then $-2ln\,[\frac{L_R}{L_F}]$ approaches 0. Thus, χ^2, the LR statistic, regardless of which two models are being compared, yields a value that lies between 0 and +∞ and has approximately a chi-square, $\chi^2(p\text{-}q)$, distribution in large samples. The degrees of freedom (*df*) for this chi-square test corresponds to *(p - q)* which is equal to the difference between the number of parameters in the two models [77, 76, 80]. The appropriate decision rule therefore is:

$$\text{if } \chi^2 \leq \chi^2{}_{(1-\alpha,p-q)}\ or\ P(\geq \chi^2) > \alpha_{cri.},\ \text{accept } H_0 \qquad (4.45a)$$

$$\text{if } \chi^2 > \chi^2{}_{(1-\alpha,p-q)}\ or\ P(\geq \chi^2) \leq \alpha_{cri.},\ \text{reject } H_0 \qquad (4.45b)$$

where the quantity $\chi^2{}_{(1-\alpha,p-q)}$ is defined to be such that $P\left(\chi^2 \geq \chi^2{}_{(1-\alpha,p-q)}\right) = \alpha$

Additionally, the restatement of the decision rule is: If the ratio L_R/L_F is small then χ^2 is too big, reject H_0, that is, $X_{q+1}, X_{q+2}\ \dots\dots\dots, X_p$ variables are highly significant. On the other hand, If $L_R = L_F$, then $-2ln\,[\frac{L_R}{L_F}]$ approaches 0. Thus, small values of G^2 lead to conclusion H_0 , that is, $X_{q+1}, X_{q+2}\ \dots\dots\dots\dots., X_p$ variables do not contribute and are nonsignificant.

4.2.7.2 The Wald Test

The Wald test can be used to assess the contribution of individual predictors or to test the significance of individual coefficient in a given model. In this test, it is interested in that the null hypothesis H_0 that the coefficient of the independent variable is equal to zero versus the alternative hypothesis H_1 that the coefficient is not zero, that is

$$H_0: \beta_j = 0\ versus\ \ H_1: \beta_j \neq 0$$

Then, the test statistic z_j is obtained by dividing *MLE* of the regression coefficient of interest by the estimate of its standard error, $se(\beta_j)$ such as

$$z_j = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \qquad (4.46)$$

This test statistic has approximately a normal (0, 1) distribution in large samples. The square of z_j is approximately a chi-square distribution with one degree of freedom. While performing the Wald test, the required information is usually provided in an output similar to Table 4.3 which lists each variable in the model followed by its ML coefficient and its standard error. RStudio package also computes the *chi-square* statistic and a *p-value*. The computed z_j can be squared and then compared with percentage points from a chi-square distribution with one degree of freedom. The p-value suggests that β_j is significantly different from zero at the 0.05, 0.01, etc. levels or not.

Table 4.3: Wald test output

Variable	Estimated Coefficient	S.E.	Chi.Sq.	P-value
X_1	$\hat{\beta}_1$	$se(\widehat{\beta}_1)$	χ^2	p
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
X_p	$\hat{\beta}_p$	$se(\widehat{\beta}_p)$	χ^2	p

The likelihood ratio and Wald test give approximately the same value in very large samples. However, in small to moderate samples, the two statistics may give very different results. The LRT is generally recommended and better than the Wald test. However, the Wald statistic is somewhat convenient to use when only one model, the full model, needs to be fit [77, 75].

4.2.7.3 Confidence Intervals (CI) on Regression Coefficients and Odds Ratios

Confidence intervals of the estimated regression coefficients in logistic regression are based on Wald statistic, z_j, in Eq. 4.46. The upper and lower points of a $100 * (1-\alpha)\%$ confidence interval(CI) for β_j can be obtained by the formula

$$\beta_j \pm z_{1-\alpha/2} * SE(\beta_j) \tag{4.47}$$

where $z_{1-\alpha/2}$ corresponds the $100 * (1-\alpha/2)$ percentage point of the normal distribution and $SE(\beta_j)$ is the estimated standard error of the β_j .

The critical values of $z_{1-\alpha/2}$ are 1.96 for α = .05 or 2.58 α = 0.01 in two tailed. For example, if we want a 95% confidence interval, then α is 0.05, 1- α/2 is 1-0.025 or 0.975, and $z_{0.975}$ percentage point which is obtained from tables of the standard normal distribution is equal to 1.96. While a confidence interval for a significant coefficient will not include zero, a confidence interval for a nonsignificant coefficient will include zero [80]. The corresponding 100 * (1 − α) percent confidence interval (CI) for the odds ratio, exp (β_j) are calculated by:

$$\exp\,[\beta_j \pm z_{1-\frac{\alpha}{2}} SE(\beta_j)] \tag{4.48}$$

We can use the confidence interval for the odds ratio to determine whether or not the odds ratio equals one. If the confidence interval does not contain one, then we conclude that the odds ratio is statistically significant [81].

4.2.8 Building Logistic Regression Models

4.2.8.1 Model Building Strategies

In simple terms, the model can be explained as a *simplified representation* of the data collected. It is a kind of mathematical equation that is used to summarise the data as closely as but also be as simple as possible. The most frequent approach to model building is to achieve the smallest model (number of variables) that still explain the data. The smallest is chosen because it is also more stable [82]. Finally, the models should be complex enough to fit the data well, but simpler models are easier to interpret [78].

4.2.8.2 Variable Selection

Variable selection is a process of reducing the number of variables in a model such that the model can be more manageable and has an interpretable set of variables [83]. Model selection procedures are also known as subset selection or variables selection procedures in the model building process. When developing a multiple regression model, we can face with a selection of many possible models. Should we include all the variables under study, or drop ones that don’t make a significant contribution to prediction? How do we decide what predictor variables to include? So, variable selection is possibly the hardest part of model building [76]. Variable selection methods aim to choose the best subset of the predictors among many variables in a given sense or to explore a set of predictors are associated with an outcome [84].

Two variable selection strategies can be applied to select variables such as purposeful variable selection algorithm which Hosmer Lemes describes [75] and automatic variable selection algorithms, which are to be explained detail below.

The variable selection process becomes more challenging as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. For example, models with several predictors often suffer from *multicollinearity* in which two or more predictor variables in a multiple regression model are highly correlated. A variable may seem to have little effect because it overlaps considerably with other predictors in the model, itself being predicted well by the other predictors. Deleting such a redundant predictor can be helpful, for instance, to reduce standard errors of other estimated effects [78].

In general, the more predictor variables included in a valid model the lower the bias of the predictions, but the higher the variance. Including too many predictors in a regression model is commonly called *over-fitting* while the opposite is called *under-fitting* [84].

4.2.8.3 Variable Selection Criteria

In most circumstances, as the number of predictors increase, the number of possible models grows rapidly. It may be impossible for an analyst to make a detailed examination of all possible regression models. If there are *p* potential predictors, then there would be 2^p possible models. For instance, when there are 10 potential *X* variables in the pool, there would be $2^{10} = 1,024$ possible regression models. With the availability of high-speed computers and efficient algorithms, running all possible regression models for 10 potential *X* variables may not be time-consuming.

There have been developed many model selection procedures to identify a small group of regression models that are “good” according to a specified criterion. With a detailed examination, they offer three to six “good” subsets according to the criteria specified, so the investigator can then carefully study these regression models for choosing the final model [76].

While many criteria for comparing the regression models have been developed in multiple linear regression models such as *R^2/ SSE, Adjusted R^2 / MSE, Mallow’s C_p Criterion, AIC / BIC, PRESS Statistic and p-values*. But for logistic regression modeling, Information criteria such as *AIC (Akaike information criterion), BIC*

(Bayesian information criterion), and p-values are often used in variable selection. For these reasons, we will focus on the use of these criteria. *AIC* and *BIC* are defined as

$$AIC = -2\ (log\text{-}\ likelihood) +\ 2p \tag{4.49}$$

$$BIC = -2\ (log\text{-}\ likelihood) + p\ ln\ (n) \tag{4.50}$$

Where $log\text{-}\ likelihood$ is the expression $l(\beta)$ in (4.26), p is the # of parameters in the model and *n* is the number of observations in the dataset.

Note that both *AIC* and *BIC* are different forms of the $-2\ (log\text{-}\ likelihood)$ and added penalties based on the number of variables such as *2p* for *AIC* and *p ln(n)* for *BIC*. Hence AIC and BIC are the measures of fit which penalize models for the number of independent variables. While adding variables to a model improves the likelihood but also increases the penalty, and the combination can result in either a better or a worse value of the criterion [83].

AIC and *BIC* criteria provide a comparison of model fit in models that are not nested. They also take into account the number of regression coefficients being tested. Even if two models have equal fit, the model having fewer predictors will have a better *AIC* fit index. A model that exhibits a good fit with a small number of predictors will have the smallest AIC values.

The *BIC* is the second measure of fit that takes into account the number of predictors. It is very much like *AIC*, however, the penalization is different. *BIC* tends to favor simpler models than *AIC*. It may be negative or positive in value; the more negative the value of the BIC, the better the fit. So, the models with small values of AIC or BIC are preferred [80].

Besides *AIC* and *BIC*, the *p-value* criterion for the many selected test static such as *t-test*, *f-test,* or *wald test* can be used for the variable selection purpose. While the *t-test* value, $t_k = \frac{\widehat{\beta}_k}{SE(\widehat{\beta}_k)}$, for testing whether or not each associated regression parameter $\beta_k = 0$ and its *p-value* are used as a decision criterion in multiple linear regression. For multiple linear regression the *t-test* value, $t_k = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$, for testing whether or not each associated regression parameter $\beta_k = 0$ and its *p-value* are used as a decision criterion. But for logistic regression, the *Wald statistic*, $w_k = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$, and its *p-value* is usually used where $\hat{\beta}_k$ is the *k*th estimated value of coefficient and $SE(\hat{\beta}_k)$ is the standard error of the coefficient [76].

The *p-value* associated with the computed *Wald statistic* can be compared to the level of significance α. In practice a level of significance of 0.05 or 0.01 is customary. The level of significance 0.05 and 0.01 are related to 95% and 99% confidence level respectively.

As a result, these criteria for comparing results might be the lowest *p-value*, lowest *AIC*, lowest *BIC*, etc.

4.2.8.4 Purposeful Variables Selection

The purposeful variable selection is an algorithm described by Hosmer-Lemeshow [75] and in which a data analyst decides at each step of the modeling process. This selection algorithm takes into account the study goals, significance tests, multicollinearity, and potential confoundings. The following several steps in abbreviated form describe the method of selecting variables to build a model [78].

***Step-1*:** At the first step in the purposeful selection, the full model that contains all variables is fitted. Then univariate analysis, Wald test, is used to identify important variables by looking at the estimated coefficients, their standard errors. The variables whose p-value < 0.25 are selected for the next step and a new reduced model is obtained.

Step-2: The reduced model selected in Step-1, containing fewer variables that are moderate significant at the 0.25 level, is now fitted, and by using the p-value of the Walt test, the importance of each variable is verified. Then, any variable that doesn't contribute to the model at the *known standard* level of significance should be eliminated, and a new model is again fitted. The new model should be compared to the old model using the LRT. If the p-value of LRT exceeds $\alpha_{cri.}$, it is concluded that the new model is not better than the old model, that is, the eliminated variable is significant.

***Step-3*:** Next, in the small model, the values of coefficients should be compared to the values of coefficients in the larger model. Let denote $\Delta\beta$ a change between reduced and large model coefficients' value for any variable can be defined as

$$\Delta\beta = 100 * \frac{\beta_{Reduced} - \beta_{Large}}{\beta_{Large}} \tag{4.51}$$

A change of coefficients ($\Delta\beta_j$) is more than 20% indicates that one or more
of the eliminated variables are important and should be added back to the model.

The process of elimination, refitting, and verifying through *Step-2* and *Step-3* is repeated until all the variables included in the model are significant. As a result, we have a model called the *main effects model,* which contains the important variables~~.~~

***Step-4*:** In the step, each continuous variable in the *preliminary main effects model,* should be checked for their linearity with the logit of the outcome. If it is not linear, the scatter plot of logit against the variable should be examined for the linearity and a suitable transformation of the variable should be found so that the logit is roughly linear.

Step-5: Once the main effects model has been obtained, the interactions among the variables in the model should be checked. A list of possible pairs of variables in the model should be created as the arithmetic product of the pairs of main effect variables.
the interactions are added to the model one by one, then its significance is identified by using the LRT.

Step-6: In this final step the interactions found significant in step 6 is added to the "main effects model" and evaluated its fit. Then, any non-significant interaction is dropped by looking at at the Wald tests and LRT for each interaction terms. In the end, we obtain a final model called preliminary final model whose overall GOF should be assessed and perform model diagnostics.

4.2.8.5 Automatic Variable Selection Algorithms

When the number of possible models, 2^p, grows rapidly with the number of predictors, evaluating all possible regression models can be a daunting task. So, a variety of automatic/ time-saving model selection methods have been developed to simplify these tasks. They can be helpful when we have many independent variables and we need some help in the investigative stages of the variable selection process. There are several available automatic model selection methods for building regression models in the literature and commercial software (e.g., RStudio, Minitab, and Stata). *Stepwise* and *all possible(best) subset* methods are two popular approaches to selecting a final set of predictor variables from a larger pool of candidate variables are methods. Most commercial software offers an option to automatically select the *best subset* or *stepwise algorithms* [76].

4.2.8.5.1 Best Subsets Algorithm

Best Subsets algorithms can select the most promising models, without having to evaluate all 2^p candidates. They require the calculation of only a small fraction of all possible regression models. These algorithms provide the best subsets according to the specified criterion and they often also identify several "good" subsets for each possible number of X variables in the model with the smallest criterion values and using much less computational effort than when all possible subsets are evaluated. They also give additional helpful information in making the final selection of the subset of X variables to be employed in the regression model. As a result, they display the best fitting models with one independent variable, two variables, three variables, and so on. The result is a display of the best fitting models of different sizes up to the full model. We need to compare the models to determine which one is the best [76].

4.2.8.5.2 Stepwise Variable Selection Algorithms

The best subset algorithms may not be feasible and require excessive computer time when the number of predictors is large (i.e., 40 or more), In such cases, *stepwise* procedures that develop the *best subsets* of X variables sequentially are generally used. An essential difference between *stepwise* procedures and the *best subsets* algorithm is that *stepwise* procedures end with the identification of a single regression model as “best.” With the *best subsets* algorithm, on the other hand, several regression models can be identified as “good” for final consideration.

The *stepwise* algorithms for multiple linear regression are easily adapted for use in logistic regression. The only change required concerns the decision rule for adding or deleting a predictor [76]. As in ordinary regression, *stepwise* logistic regression algorithms can select or delete predictors from a model in a stepwise manner. There are three common related stepwise approaches for doing this, such as *forward selection, backward elimination,* and *stepwise selection* based on a chosen criterion.

The *backward elimination* algorithm begins with a complex model and the model improves step by step by dropping a variable from the model at each step according to a criterion (e.g. min *AIC* or *BIC*, max *p-value*, *p-value* greater than α_{crit}). The process stops if another step does not show a further improvement of the model.

The forward selection algorithm builds the model starting with no variables in the model and adds useful variables one by one. It tests the addition of each variable not in the model-based a chosen criterion (e.g.min *AIC*, min *BIC*, min *p-value*, or *p-value* less than α_{crit}). The process stops if another step does not show a further improvement of the model.

The stepwise selection approach combines both forward selection and backward elimination. It tests at each step for variables to be added OR removed according to criterion [84].

These *stepwise* algorithms based on *p-values* have some advantages such as easy to explain, easy to compute, and widely used. But they have also some disadvantages. Because when we drop and add variables one at a time, it is possible to miss the 'optimal model' or method may overstate the significance of results. So we should not trust the *p-values* too much. They can assist in building a model but they do have some drawbacks so they have to be used with caution and the final model has always to be reviewed by the researcher [78].

Additionally, the identification of a single regression model as "best" by the stepwise procedures is a major weakness of these procedures. Experience has shown that each of the stepwise procedures can sometimes wrongly identify a suboptimal regression model as "best." Besides, the identification of a single regression model may hide the fact that several other regression models may also be "good." [76].

Finally, it should be noted that there is not a correct model and every model is a simplification of reality. But models can explain reality well to different degrees and they can provide insight into relationships between predictors and response [78].

4.2.9 Working with Categorical Variables

In epidemiologic studies, continuous variables are generally categorized into quartiles or quintiles to illustrate the relationship between continuous exposure and a binary outcome [85]. The odd ratios for successively higher quartiles and IQR can be used to describe a relationship between an exposure and an outcome by three or four separate estimates by using the lowest quartile as the reference category.

In a regression analysis, R identifies categorical variables as ordered or nonordered factors and p-1 dummy variables if a categorical variable has p category(level). So, the regression is implemented with p-1 dummy variables instead of one categorical variable having p level. When coding categorical variables, dummy coding is probably the most commonly used one. It compares each level of the categorical variable to a fixed reference level [81]. In our study, categorical variables are divided into four levels coded as 1, 2, 3, or 4 to show quarters (*Q1, Q2, Q3, or Q4*). Here, the dummy variable showing the reference(smallest) levels of a categorical variable is omitted and three dummy variables are constructed to represent the levels of the categorical variable encoding the. For example, the dummy variables for NO_2 quartile levels can be expressed as shown in Table 4.4.

Table 4.4: The dummy variables for NOX Qurtile levels

	Dummy Variables		
NO_2 Quartile Levels (1. Year)	**D_1**	**D_2**	**D_3**
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	0

The dummy variables D_1, D_2, and D_3 take 1, 0, and 0 values respectively to show the 2nd quartile level of the NO_2. The log-odds model by using Eq. 4.16 can be expressed as with the dummy variable as

$$logit(\pi) = log(odds) = (\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3) \tag{4.52}$$

The log-odds for different levels can be expressed as in Table 4.5.

Table 4.5: The Log-odds of Disease and Odds of Disease for Quartile levels

Quartile Levels	**Log-odds of Disease**	**Odds of Disease**
1	$(\beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0)) = \beta_0$	$\mathrm{exp}(\beta_0)$
2	$(\beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0)) = \beta_0 + \beta_1$	$\mathrm{exp}(\beta_0 + \beta_1)$
3	$(\beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(0)) = \beta_0 + \beta_2$	$exp(\beta_0 + \beta_2)$
4	$(\beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(1)) = \beta_0 + \beta_3$	$exp(\beta_0 + \beta_3)$

Thus, the odds ratio for second quartile level, odds being in the *2nd* quartile(*Q2*) versus odds being first quartile(*Q1*), that is, going from quartile = 1 to quartile = 2 is calculated by using Eq. 4.40 as

$$\mathrm{OR} = \frac{odds_{\mathrm{Q2}}}{odds_{\mathrm{Q1}}} = \frac{\mathrm{e}^{\beta_0+\beta_1}}{\mathrm{e}^{\beta_0}} == \mathrm{e}^{\beta_1} \tag{4.53}$$

Hence, e^{β_1} is the relative increase in the odds of disease, going from quartile = 1 to quartile = 2.

4.2.10 Overall Model Evaluation / Goodness of Fit (GOF)

After obtaining a final model containing needed variables, we need to know how well the model fits with the data by comparing the observed and predicted logits or probabilities for all predictors. This is referred to as a GOF [77]. Residuals can be used in testing the GOF of the model. Several types of residuals play an important role in the analysis of logistic regression to provide useful information about the model and to examine the fit of the logistic model as in multiple linear regression. So, the assessment of the fit of a GLM typically begins with looking at the residual deviance and Pearson residuals for the model. They are used in Pearson's chi-squared and deviance GOF tests to assess model adequacy [86, 87].

4.2.10.1 Pearson Chi-squared GOF Test

In linear regression, the method of least squares analysis is based on the total sum of squared residuals. The total variation, $(Y_i - \bar{Y})$, in the response variable, Y, can be subdivided into two sums of squares components:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (4.54)$$

where Y_i denotes the ith observed value and $\bar{Y}$ denotes the $mean(Y)$ under the model.

Then the total variation symbolically can be written as

$$SST = SSR + SSE \qquad (4.55)$$

where *SST* (the total sum of squares) is the total variation in the response variable, *SSR*(regression sum of squares) is residual variation, or variation between fitted value and mean of the fitted values, and *SSE*(error sum of squares) is variation between observation Y_i and fitted value [76].

So, R^2 becomes standard GOF measurement tools for linear Regression. It is equal to the ratio of the variance of the fitted values to the total variance ranging from 0 to 1. The relationship between R^2 and these variations can be defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{1}^{n}(Y_i - \bar{Y})^2}$$

While the standard GOF measure for linear regression is the R^2 statistic, it is not suitable for use with the logistic model. The ones designed for logistic models, which are called *Pseudo-R2* statistics, have been generally unsuccessful [88].

In logistic regression, residuals can be used to evaluate the GOF and to measure the difference between the observed and fitted values. For the following logistic regression model,

$$logit(\pi_i(x_i)) = (\beta_0 + \beta_1 X_{i1} +. \beta_2 X_{i2} \ldots\ldots\ldots\ldots\ldots\ldots.. + \beta_p X_{ip}) \quad (4.56)$$

The raw residual for each observation in the model can be defined as

$$Y_i - \hat{Y}_i \quad (4.57)$$

where Y_i is the observed and $\hat{Y}_i$ is the fitted value for the ith subject.

They should be close to each other for a better fit. Most of the residuals are based on the raw residual. Pearson residual is one of them used in logistic regression to examine the fit of the logistic model. Itis simply the raw residual which is divided by its estimated standard error, $se(Y_i - \hat{Y}_i)$, is defined as:

$$r_i = \frac{Y_i - \hat{Y}_i}{se(Y_i - \hat{Y}_i)} \quad (4.58)$$

Since the standard deviation of the binomial distribution is $\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}$, Pearson residual can be adjusted for the binary models as

$$r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}} \quad (4.59)$$

Then, a Pearson test statistic following χ^2, or *chi-squared*, distribution with n - (k + 1) degrees of freedom can be formed based on this residual in Eq. 4.59 by summing the squares of them as

$$\chi^2 = \sum_{i=1}^{n} {r_i}^2 \quad (4.60)$$

where n = the number of samples, k = the number of predictors in the model so that *p-values* can be calculated [88].

Finally, another easy form to remember of the Pearson GOF test is

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{4.61}$$

where O_i is the number of observed items in the ith category and E_i is the expected frequency, which is the number of expected items in that ith category [86, 75, 78, 89].

4.2.10.2 Deviance & Goodness of Fit (GOF)

For the logistic regression model deviance a widely used GOF measure which shows how well the fitted model fits the raw data [77]. Deviance, *D*, which is the sum of the differences between the saturated and proposed model log-likelihoods is defined as:

$$D = -2\sum[LL(y_i; y_i) - LL(y_i; \hat{\pi}_\iota)] \tag{4.62}$$

where y_i is the response, $\hat{\pi}_\iota$ is the fitted value.

If the model has a good fit, the deviance will be small. Otherwise, it will be a high and bad fit. By using Eq. 4.23, the log-likelihood function can be achieved for each observation in the regression model as:

$$LL(y_i, \hat{\pi}_\iota) = \sum[y_i * ln(\frac{\hat{\pi}_\iota}{1-\hat{\pi}_\iota})] + ln(1-\hat{\pi}_\iota)] \tag{4.63}$$

On the other hand, the saturated log-likelihood, LL_S, is calculated by substituting y_i for every $\hat{\pi}_\iota$ in the logistic log-likelihood function as:

$$LL_S(y_i, \hat{\pi}_i) = \sum[y_i * ln(\frac{y_i}{1-y_i})] + ln(1-y_i)] \tag{4.63}$$

Then, by using Eq. 4.62, the deviance is calculated as:

$$D = -2\sum[y_i * ln(\frac{\hat{\pi}_i}{1-\hat{\pi}_i})] + ln(1-\hat{\pi}_i)] \; - [y_i * ln(\frac{\hat{\pi}_i}{1-\hat{\pi}_i})] + ln(1-\hat{\pi}_i)] \tag{4.64}$$

Additionally, for binary models deviance in a simple form is as follows:

$$D = -2\sum[y_i * ln(\frac{y_i}{\hat{\pi}_i}) + (1-y_i) * \; ln\,(\frac{(1-y_i)}{(1-\hat{\pi}_i)})] \tag{4.65}$$

Then, the deviance residual, which represent the contributions of individual samples to the deviance *D*, is defined as:

$$r_i \ = \ sgn(y_i \ - \ \hat{\pi}_i) \ * \ sqrt(deviance) \qquad (4.66)$$

Again, as in Pearson GOF test, the sum of squared *deviance residuals* produces the *deviance* GOF [88]:

$$D = \sum {r_i}^2 \qquad (4.67)$$

Although the Pearson GOF test was more popular, now the deviance residual is generally preferred over the Pearson residual [88].

After all, in various texts, the deviance can be defined as minus twice the value of the log of the LR and is abbreviated as -*2LL*=-2 ln(likelihood ratio) [80].
The deviance, *D,* for any regression model is defined as

$$D \ = -2ln\left[\frac{max.\ likelihood\ of\ proposed\ model}{\text{max.}\ likelihood\ of\ saturated\ model}\right] \qquad (4.68)$$

Deviance compares any proposed model to the saturated model where the number of parameters equals the number of observations to determine how well the proposed model fits the data [80, 77].

For instance, let L_P denote the max. likelihood value for a proposed model with a few predictors and L_S denote the likelihood value for the most complex model possible or saturated model. If we compare a model having a few predictors with the most complex model, then the deviance is:

$$D = -2\ ln(\frac{L_p}{L_S}) = -2\ ln\ L_P - (-2\ L_S) \qquad (4.69)$$

In particular, if $L_P = L_S$, then the deviance $D = -2\ ln(\frac{L_P}{L_S}) = -2\ ln(1) = 0$. In contrast, if $L_P \ll L_S$, then the ratio $\frac{L_P}{L_S}$ is a small fraction, so that $ln(\frac{L_P}{L_S})$ is a large negative number and $-2\ ln(\frac{L_P}{L_S})$ will be a large positive number. Thus, the deviance *D* values range from zero to larger and larger positive numbers and follow a chi-squared distribution whose degrees of freedom are equal to the difference in the number of parameters between the saturated and proposed models: *p – q* [77].

Finally, if the deviance value is relatively small for a proposed model when two likelihoods are close to each other, then it is the best model. Otherwise, if the value of the deviance is large for a proposed model then it is the worse model; that is, deviances are measures of “badness of fit.” [80].

4.2.10.3 Model Comparison Using Deviance

The nested models can be compared by comparing their deviances,i.e., the difference in deviances can be used to compare them. The difference in deviances between the two models is equivalent to the LRT. It can be shown with the following example: consider two nested models, denoted by M_0 and M_1, such that M_0 is a special case of M_1,i.e., M_0 is a reduced model containing less predictor than a more complex model M_1. Additionally M_S is a saturated model, that is, the most complex model possible. Using Expression (4.69), deviances D_0 and D_1 for reduced and more complex model respectively can be written as:

$$D_0 = -2\,ln(\tfrac{L_0}{L_S}) = -2\,ln\,L_0 - (-2\,L_S) \tag{4.70}$$

$$D_1 = -2\,ln(\tfrac{L_1}{L_S}) = -2\,ln\,L_1 - (-2\,L_S) \tag{4.71}$$

where L_S is the max. likelihood value for the saturated model. By taking the difference between deviances in two models:

$$D_0 - D_1 = -2\,ln\,L_0 - (-2\,L_s) - [-2\,ln\,L_1 - (-2\,L_S)] \tag{4.72}$$

$$= -2\,ln\,L_0 - 2\,ln\,L_1 = -2\,ln(\frac{L_0}{L_1})$$

$$= LR \tag{4.73}$$

Then, we have shown that the difference in deviances between the two models is equal to the LR. Then we can compare the models by comparing their deviances. If $L_0 < L_1$, the difference in deviances, $D_0 - D_1$, becomes a large positive number. Then, we conclude that reduced model, M_0, fits poorly compared with the more complex model, M_1. For large samples, the statistic has an approximate chi-squared distribution, with *df* equal to the difference between the residual *df* values for the separate models. This *df* value equals the number of additional parameters that are in M_1 but not in M_0. Large test statistics and small *p-values* suggest that model M_0 fits more poorly than M_1 [78].

4.2.10.4 Hosmer-Lemeshow GOF Test

Hosmer and Lemeshow have developed a commonly used test to assess the GOF for binary logistic models. [74]. Here are the summarized steps in [77] to compute the Hosmer-Lemeshow test. Firstly, the predicted probabilities for all observations in the dataset are computed and sorted in descending order. Next, the ordered predicted probabilities of the model are divided into groups or quartiles. Each group is a range of probabilities. The number of groups is generally chosen as 10. After that, the observed and expected number of 0s and 1s in each group are calculated and are compared to each other. Finally, a Pearson Chi-squared test, which sums the difference between predicted and observed frequencies and compares them, is performed with the following formula:

$$H = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{E_g} \tag{4.74}$$

where O_g and E_g denote the number of observed and expected cases in the *j*th group.

The test statistic asymptotically follows a χ^2 distribution. It returns as an output a χ^2 value and a *p-value* ($P(\geq \chi^2$). The appropriate decision rule to interpret the output therefore is

$$\text{if } \chi^2 \leq \chi^2_{(1-\alpha,\text{df})} \; or \; P(\geq \chi^2\,) > \alpha_{cri}\,, \text{ fail to reject(accept) } H_0 \tag{4.75}$$

$$\text{if } \chi^2 > \chi^2_{(1-\alpha,\text{df})} \; or \; P(\geq \chi^2) \leq \alpha_{cri.}\,, \text{ reject } H_0 \tag{4.76}$$

where the quantity $\chi^2_{(1-\alpha,\text{df})}$ is defined to be such that $P\left(\chi^2 \geq \chi^2_{(1-\alpha,\text{p}-\text{q})}\right) = \alpha$
and *df* is degrees of freedom of the model.

While smaller χ^2 values with large *p-value*s greater than 0.05, closer to 1, indicate a well-fitted model, larger χ^2 values with $p < 0.05$ indicate a poor fit to the data where H_0: observed-predicted=0, the model fits vs. H_1: observed-predicted≠0, the model does not fit. Hosmer and Lemeshow do not recommend the use of this test when there is a small number of objects, less than 400 [77, 90].

4.3 Used Softwares

4.3.1 Rstudio IDE

Rstudio is a software application that provides many desired features and makes it easier programming with R. It includes an editor that supports direct code execution and tools for plotting [91].

4.3.1.1 What is R?

"R is a language and environment for statistical computing and graphics. *R* provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed." [92].

4.3.1.2 Visualising Data

While working in data mining projects, graphical tools allow us to examine the characteristics of data visually, see the distributions of the variables. R provides many options to present data and allows us to program the visualizations such as histogram and density or scatter plot [93].

4.3.1.2.1 Histogram

A histogram is a useful graphical tool to display the frequency of the data intervals and the distribution of the data quickly. Also, an idea about the skewness of the data can be achieved [93].

Histograms can be created with the function *hist(x)* in R where *x* is a variable to be viewed [94].

For example, Avg. NO_2 during the third trimester in the “Exposure” table has been partitioned into ranges, and the frequency of each range is displayed as the bar [93].

The results of the following R-code for the histograms were plotted in Figure 4.7

Histogram Codes

```
#1- Histogram with the option freq=FALSE creates a plot based on
frequencies
hist(exposure$no2_Trim3, -
 freq=TRUE,  breaks=12,
     col="red",
     xlab="Avg. NO2 in ppb",
     main="Histogram of Avg. NO2 during the Third trimester")

#2-Histogram with the option freq=TRUE creates a plot based on probability
densities
hist(exposure$no2_Trim3,
    freq=FALSE,
    breaks=12,
    col="red",
    xlab="Avg. NO2 in ppb",
    main="Histogram of Avg. NO2 during the Third trimester")
```

We might observe that the most frequent range of values is in the 7-8 partition.

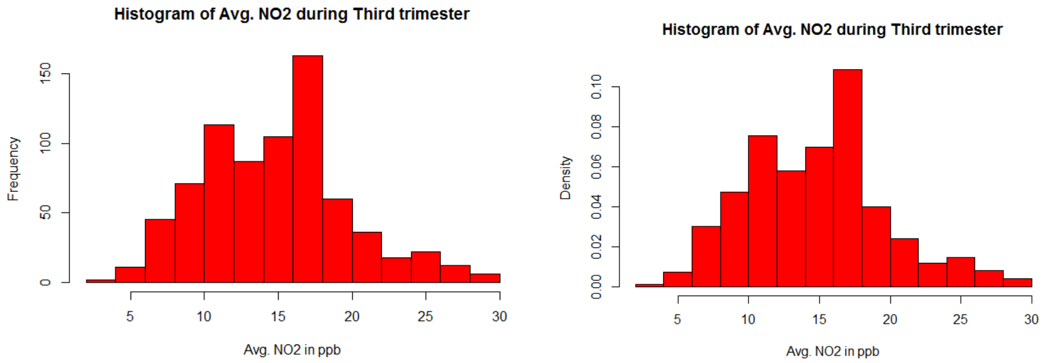


Figure 4.7: Histogram Examples

4.3.1.2.2 Density Plots

A *density plot can be thought of* as a “continuous histogram” of a variable to examine the distribution of a numerical variable. It is a more understandable display of the actual data [93].

The results of two density examples given in the next following code are plotted in Figure 4.8.

Density plot Codes

```
# Density plot that’s not being superimposed on another graph
par(mfrow=c(2,1))
d <- density(exposure$no2_Trim3)
plot(d, main="Density of Avg. NO2 during the Third trimester")

# Density plot that’s  being superimposed on another graph
hist(exposure$no2_Trim3,
     freq=FALSE,
     breaks=12,
     col="red",
     xlab="Avg. NO2 in ppb",
     main="Histogram of Avg. NO2 during the Third trimester")
lines(density(exposure$no2_Trim3), col="blue", lwd=2)
```

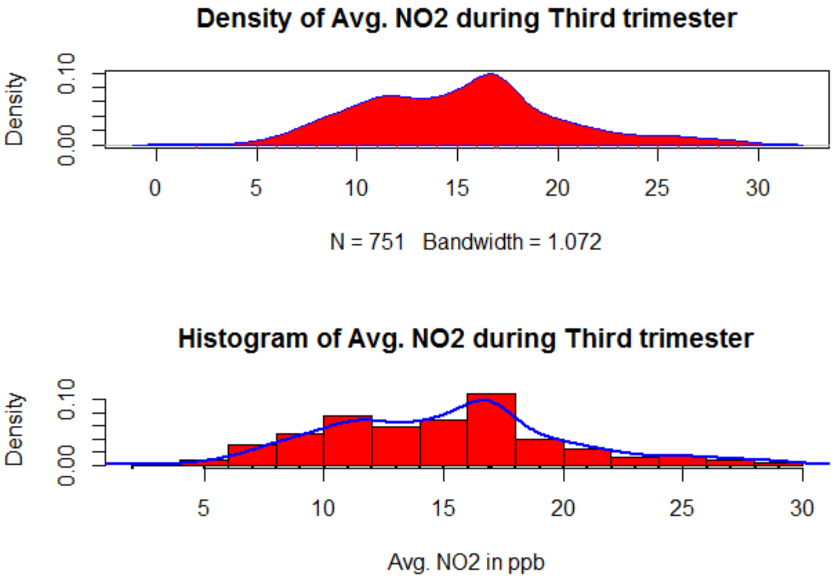


Figure 4.8: Density Plot Examples

4.3.2 Microsoft Access

Microsoft Access is a database application with a graphical user interface and software development tools. Tables, queries, forms, reports, and macros can be created in its interface easily.

4.3.3 Minitab

Minitab is a software product that helps you to analyze the data [95]. It consists of two main parts which are the data and the session windows. When you start Minitab, they are displayed by default. The data window is where the data are entered. The session window is where commands and reports are displayed. A lot of the complex analysis can be done by Minitab through Minitab's menus or Minitab macros [96].

5. DATA ANALYZING

5.1 Data Preprocessing

Microsoft Access, RStudio, and Minitab software were used while preprocessing. Firstly, the Microsoft Access tool was used to achieve the final table. As explained in chapter 3, all the datasets in Table 3.1 were imported to new tables in Microsoft Access. Next, these tables were also transformed and transposed into new temporary tables such as tempNO2, tempO3, tempPM10, and temp25 respectively with a large number of rows but a small number of dimensions as in Table 3.8 and Table 3.8. Then, the “Exposure” table was created by joining tempNO2, tempO3, tempPM10, and temp25 tables with the “subject_key” field and inserting their data into final “Exposure” table shown in Table 3.10.

When the “Exposure” table was examined then the phenotype of the 152 subjects was labeled as "Not defined", so they were excluded by deleting. Additionally, 136 subjects whose phenotypes labeled as "NEUROLOGICAL CONTROL" were also deleted.

Then my thesis data, “Exposure” table, was formed with 751 subjects shown in Table 5.1, 119 of them are females and other 632 are males, that is, 15.85 percent of subjects are female and 84.15 is male shown in Table 5.2. Finally, 63.78% of subjects are Autistic and 36.22% are not shown in Table 5.3.

Table 5.1: The summary of “Exposure” table based on phenotype

Phenotype	# of Subjects
AUTISM SPECTRUM AFFECTED	141
AUTISM SPECTRUM SEVERELY AFFECTED	338
TYPICAL CONTROL	272
Total	751

Table 5.2: The summary of "Exposure" table based on gender

Gender	# of Subjects	%
FEMALE	119	15.85%
MALE	632	84.15%

Table 5.3: The summary of "Exposure" table based on Autistic

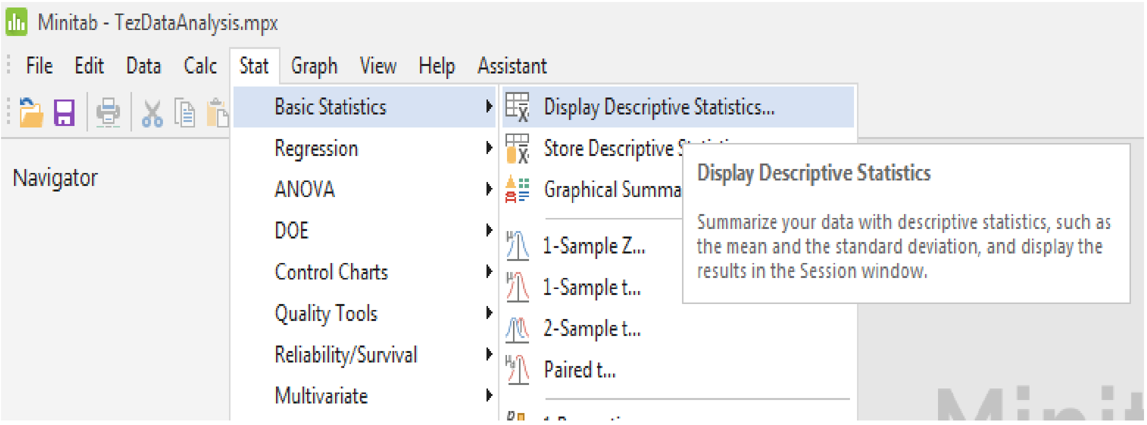
Autistic	# of Subjects	%
Y	479	63.78%
N	272	36.22%
Total	751	

Next, the "Exposure" table was exported as a comma "," delimited text file, "NdarExposure.txt", from Microsoft Access. Then, it was loaded into RStudio from the *Ndarexposure.txt* file with the following R command.

```
>exposure <- read.csv("QNdarExposure.txt", header=TRUE,dec=".")
```

5.1.1 Some Descriptive Statics

The "Exposure" table was also imported into Minitab. Then, the statistical properties of the dataset were obtained by using the commands as in Figure 5.1



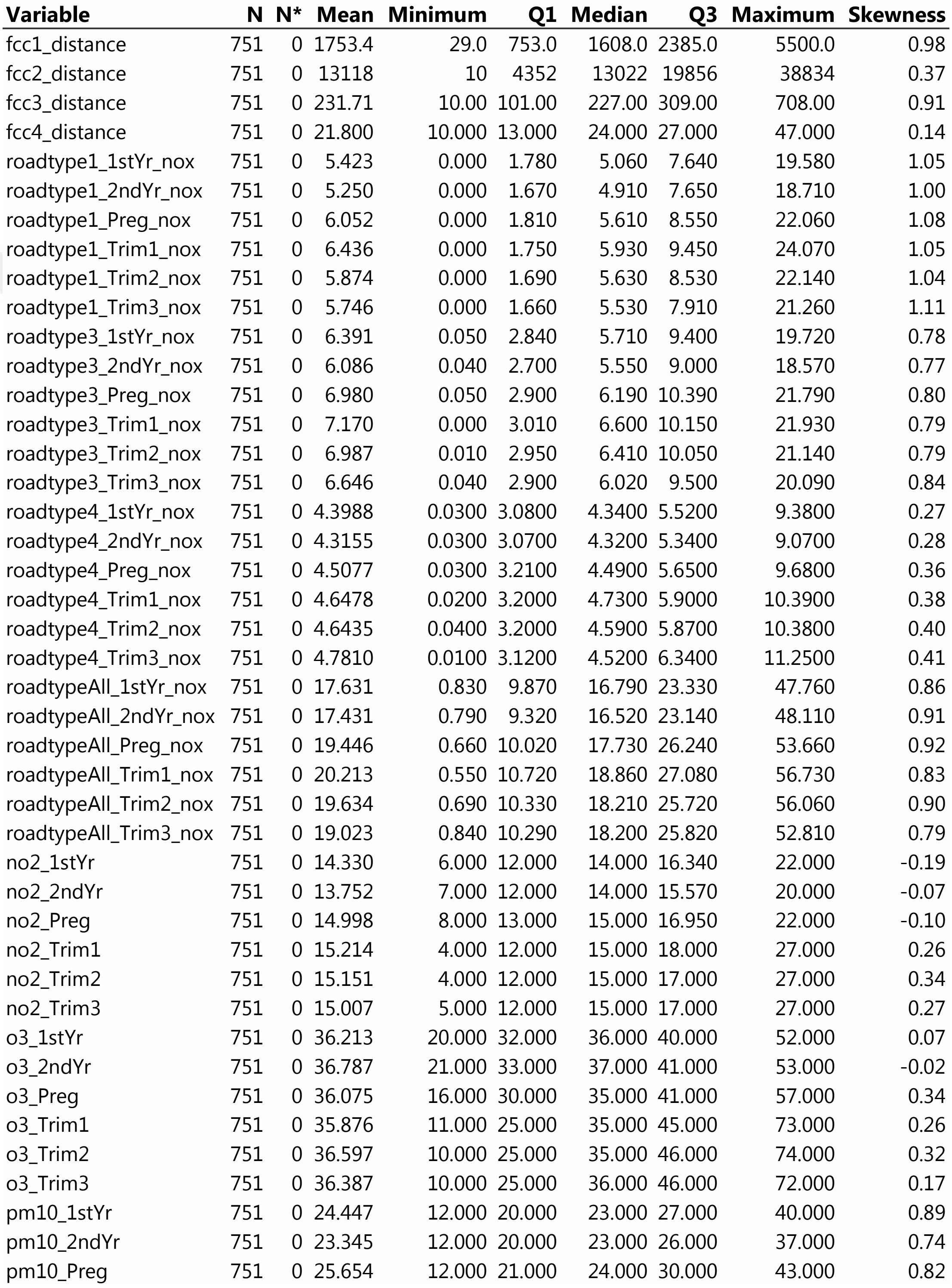
Summarize your data with descriptive statistics, such as the mean and the standard deviation, and display the results in the Session window.

Figure 5.1: Commands to achieve Statistical measures for "NdarExposure.txt"

Statistical measures for "NdarExposure.txt" were attained shown in Table 5.4. They show the data summarization (*mean and median), the* central tendency of data (minimum, maximum, first, and third *quartiles), and* the dispersion of data (*skewness).* It also includes NA's (null values) information.

roadtype2_1stYr_nox,roadtype2_2ndYr_nox,roadtype2_Preg_nox,roadtype2_Trim1_nox, roadtype2_Trim2_nox and roadtype2_Trim3_nox atttributes have nearly 500/751 zero values, so they were excluded from working dataset.

Table 5.4: Statistical measures for "NdarExposure.txt"

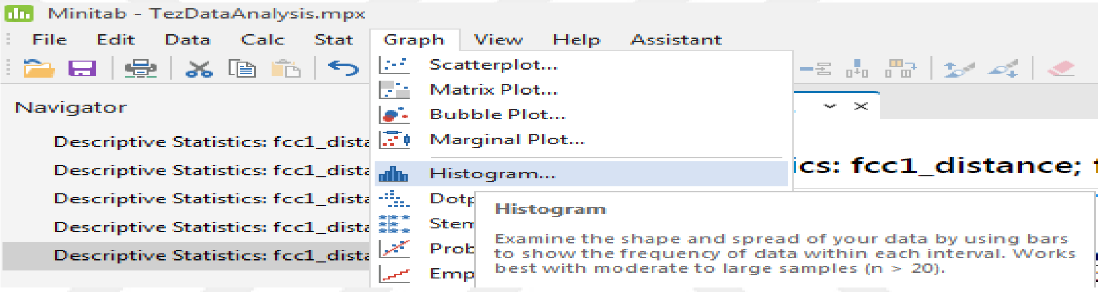


Variable	N	N*	Mean	Minimum	Q1	Median	Q3	Maximum	Skewness
fcc1_distance	751	0	1753.4	29.0	753.0	1608.0	2385.0	5500.0	0.98
fcc2_distance	751	0	13118	10	4352	13022	19856	38834	0.37
fcc3_distance	751	0	231.71	10.00	101.00	227.00	309.00	708.00	0.91
fcc4_distance	751	0	21.800	10.000	13.000	24.000	27.000	47.000	0.14
roadtype1_1stYr_nox	751	0	5.423	0.000	1.780	5.060	7.640	19.580	1.05
roadtype1_2ndYr_nox	751	0	5.250	0.000	1.670	4.910	7.650	18.710	1.00
roadtype1_Preg_nox	751	0	6.052	0.000	1.810	5.610	8.550	22.060	1.08
roadtype1_Trim1_nox	751	0	6.436	0.000	1.750	5.930	9.450	24.070	1.05
roadtype1_Trim2_nox	751	0	5.874	0.000	1.690	5.630	8.530	22.140	1.04
roadtype1_Trim3_nox	751	0	5.746	0.000	1.660	5.530	7.910	21.260	1.11
roadtype3_1stYr_nox	751	0	6.391	0.050	2.840	5.710	9.400	19.720	0.78
roadtype3_2ndYr_nox	751	0	6.086	0.040	2.700	5.550	9.000	18.570	0.77
roadtype3_Preg_nox	751	0	6.980	0.050	2.900	6.190	10.390	21.790	0.80
roadtype3_Trim1_nox	751	0	7.170	0.000	3.010	6.600	10.150	21.930	0.79
roadtype3_Trim2_nox	751	0	6.987	0.010	2.950	6.410	10.050	21.140	0.79
roadtype3_Trim3_nox	751	0	6.646	0.040	2.900	6.020	9.500	20.090	0.84
roadtype4_1stYr_nox	751	0	4.3988	0.0300	3.0800	4.3400	5.5200	9.3800	0.27
roadtype4_2ndYr_nox	751	0	4.3155	0.0300	3.0700	4.3200	5.3400	9.0700	0.28
roadtype4_Preg_nox	751	0	4.5077	0.0300	3.2100	4.4900	5.6500	9.6800	0.36
roadtype4_Trim1_nox	751	0	4.6478	0.0200	3.2000	4.7300	5.9000	10.3900	0.38
roadtype4_Trim2_nox	751	0	4.6435	0.0400	3.2000	4.5900	5.8700	10.3800	0.40
roadtype4_Trim3_nox	751	0	4.7810	0.0100	3.1200	4.5200	6.3400	11.2500	0.41
roadtypeAll_1stYr_nox	751	0	17.631	0.830	9.870	16.790	23.330	47.760	0.86
roadtypeAll_2ndYr_nox	751	0	17.431	0.790	9.320	16.520	23.140	48.110	0.91
roadtypeAll_Preg_nox	751	0	19.446	0.660	10.020	17.730	26.240	53.660	0.92
roadtypeAll_Trim1_nox	751	0	20.213	0.550	10.720	18.860	27.080	56.730	0.83
roadtypeAll_Trim2_nox	751	0	19.634	0.690	10.330	18.210	25.720	56.060	0.90
roadtypeAll_Trim3_nox	751	0	19.023	0.840	10.290	18.200	25.820	52.810	0.79
no2_1stYr	751	0	14.330	6.000	12.000	14.000	16.340	22.000	-0.19
no2_2ndYr	751	0	13.752	7.000	12.000	14.000	15.570	20.000	-0.07
no2_Preg	751	0	14.998	8.000	13.000	15.000	16.950	22.000	-0.10
no2_Trim1	751	0	15.214	4.000	12.000	15.000	18.000	27.000	0.26
no2_Trim2	751	0	15.151	4.000	12.000	15.000	17.000	27.000	0.34
no2_Trim3	751	0	15.007	5.000	12.000	15.000	17.000	27.000	0.27
o3_1stYr	751	0	36.213	20.000	32.000	36.000	40.000	52.000	0.07
o3_2ndYr	751	0	36.787	21.000	33.000	37.000	41.000	53.000	-0.02
o3_Preg	751	0	36.075	16.000	30.000	35.000	41.000	57.000	0.34
o3_Trim1	751	0	35.876	11.000	25.000	35.000	45.000	73.000	0.26
o3_Trim2	751	0	36.597	10.000	25.000	35.000	46.000	74.000	0.32
o3_Trim3	751	0	36.387	10.000	25.000	36.000	46.000	72.000	0.17
pm10_1stYr	751	0	24.447	12.000	20.000	23.000	27.000	40.000	0.89
pm10_2ndYr	751	0	23.345	12.000	20.000	23.000	26.000	37.000	0.74
pm10_Preg	751	0	25.654	12.000	21.000	24.000	30.000	43.000	0.82

pm10_Trim1	751	0	25.227	9.000	20.000	24.000	30.000	45.000	0.64
pm10_Trim2	751	0	25.053	11.000	19.000	23.000	30.000	46.000	0.72
pm10_Trim3	751	0	24.929	8.000	19.000	23.000	30.000	46.000	0.65
pm25_1stYr	751	0	12.289	7.000	11.000	12.000	13.630	18.000	0.11
pm25_2ndYr	751	0	11.777	6.000	10.000	12.000	13.190	17.000	-0.05
pm25_Preg	751	0	13.673	4.000	11.000	13.000	16.000	23.000	0.52
pm25_Trim1	751	0	13.471	4.000	9.000	13.000	17.000	29.000	0.74
pm25_Trim2	751	0	13.366	4.000	9.000	12.000	17.000	29.000	0.75
pm25_Trim3	751	0	13.848	3.000	8.000	12.000	18.000	33.000	0.89

5.1.2 Visualization of the Distributions

The Histograms of the dataset were obtained by using the commands as in the Figure 5.2


Minitab - TezDataAnalysis.mpx
File
Edit
Data
Calc
Stat
Graph
View
Help
Assistant
Navigator
Descriptive Statistics: fcc1_dista
Descriptive Statistics: fcc1_dista
Descriptive Statistics: fcc1_dista
Descriptive Statistics: fcc1_dista
Descriptive Statistics: fcc1_dista
Scatterplot...
Matrix Plot...
Bubble Plot...
Marginal Plot...
Histogram...
Dotp
Sten
Prob
Emp
ics: fcc1_distance; f
Histogram

Examine the shape and spread of your data by using bars to show the frequency of data within each interval. Works best with moderate to large samples (n > 20).

Figure 5.2: Commands to achieve Histograms for “NdarExposure.txt”

Histogram command is used to graphically summarize the distribution (spreads, skewness) of the dataset as shown in Figure 5.3 and Figure 5.4.

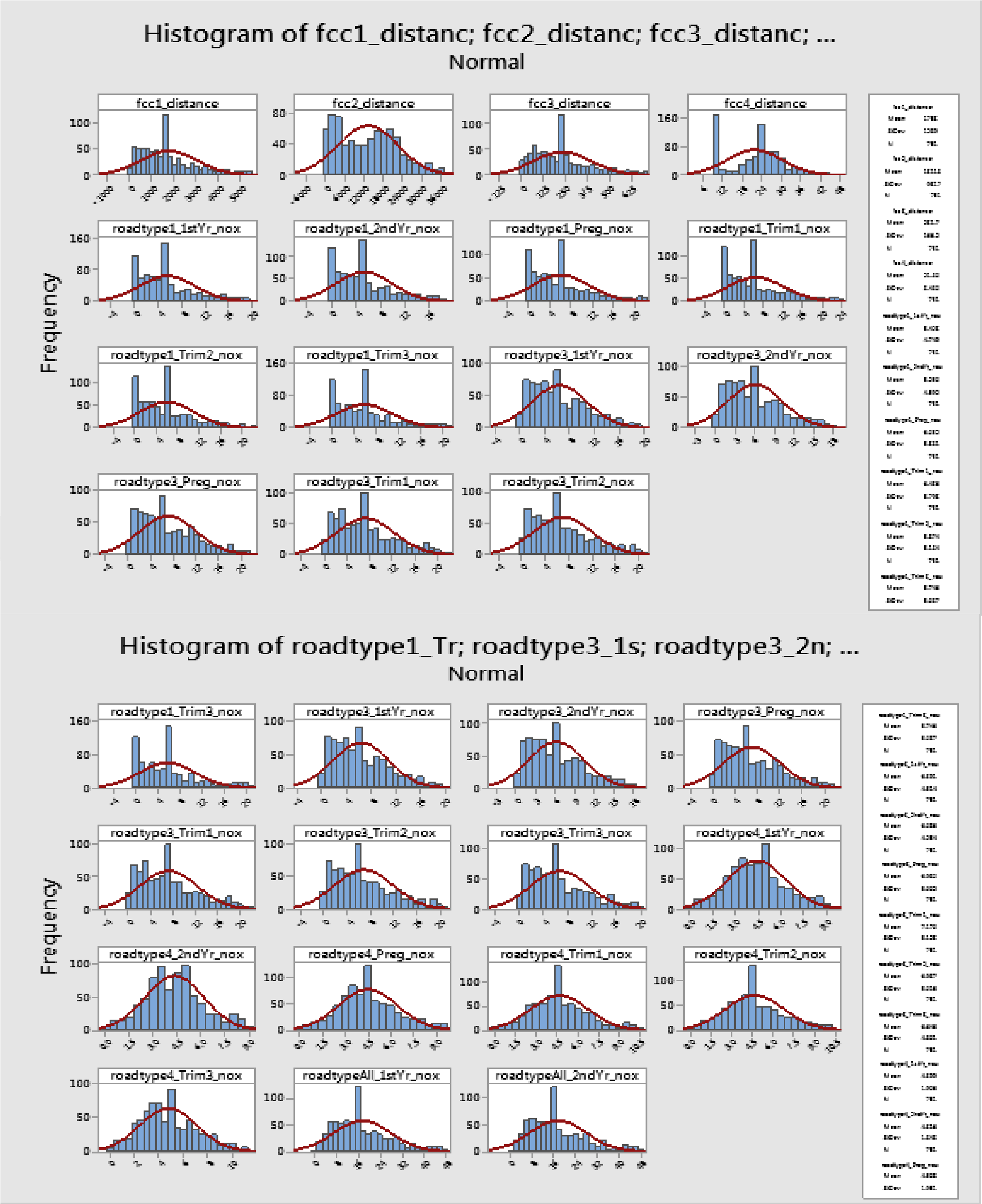


Figure 5.3: Histogram of some attributes(a)

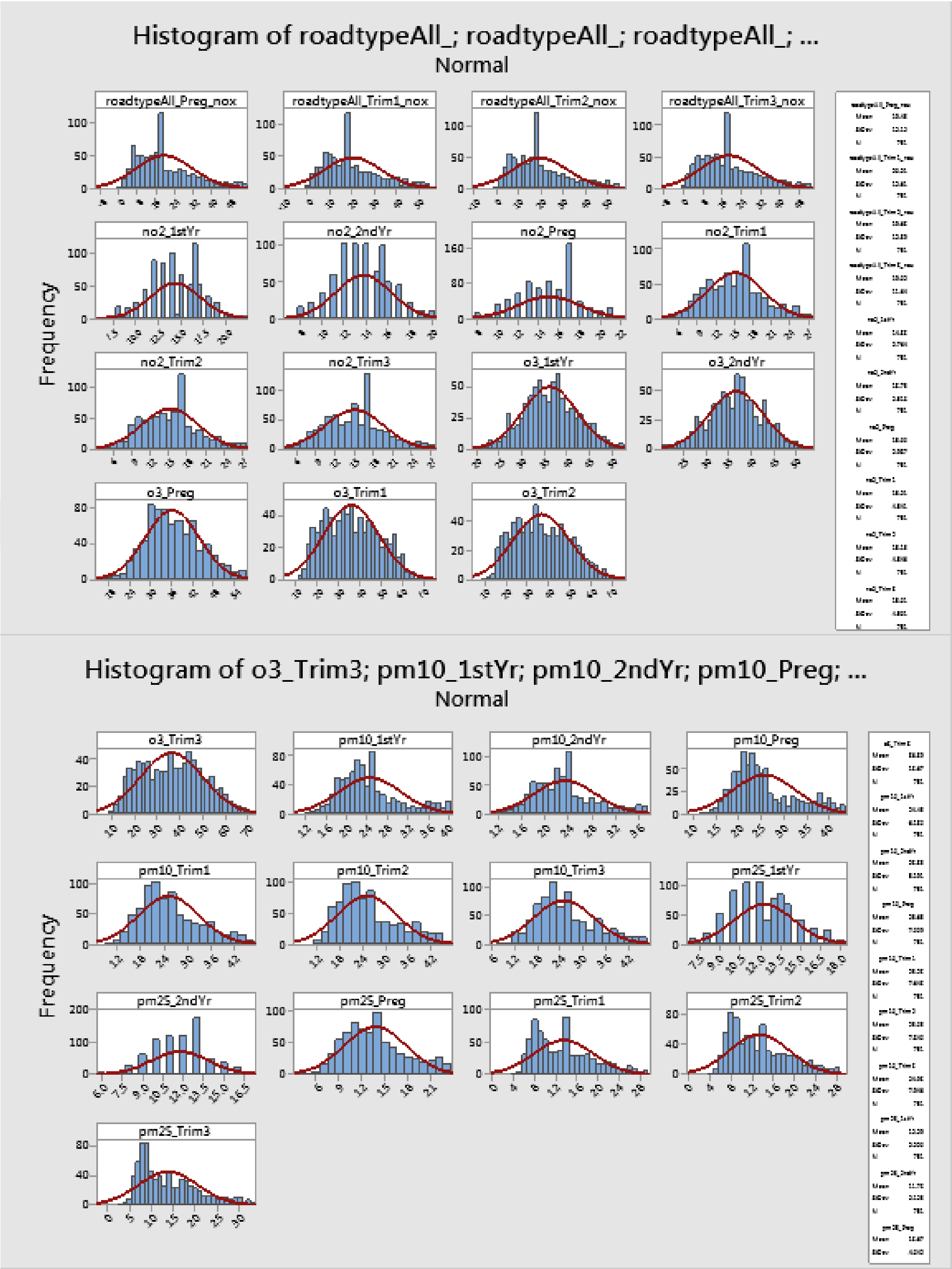


Figure 5.4: Histogram of some attributes(b)

5.2 Missing Values

To fill in the missing value, a measure of central tendency for the attribute (e.g., the mean or median) is used by using Histogram information. The mean was used for normal (symmetrical) data distributions, while the median was used for skewed data distribution. The missing values in the "Exposure" table were replaced with the following code in Rstudio.

Replacing Missing Values Codes

```
#Replacing Missing Values In R with column mean/medium
for(i in 1:67){
cn=c("fcc1_distance","fcc3_distance","fcc4_distance",.....)
t=which(cn==colnames(exposure)[i])
 if (length(t)>0) ftype=1 else  ftype=2
 if (ftype==1)
exposure[is.na(exposure[,i]), i] <- median(exposure[,i], na.rm = TRUE)
 else
  exposure[is.na(exposure[,i]), i] <- mean(exposure[,i], na.rm = TRUE)
}
write.table(exposure, file = "Ndarexposure.txt", sep = ",", row.names =
FALSE,col.names = TRUE )
```

5.3 Outliers

Outliers are extreme observation values that appear to be different from the remaining data in the existing dataset. They are the values that are either too large or too small. Outliers in "Exposure" table were detected by using *Box Plot Rule* and replaced with median/medium with the following code for "Exposure" table in RStudio

Remove outliers Codes

```
#remove outliers
for(i in 1:67){
    t=which(cn==colnames(exposure)[i])
   if (length(t)>0) ftype=1 else  ftype=2
    y=remove_outliers(exposure[,i],filltype=ftype)
    exposure[,i]=y
  }
#*****remove_outliers function
remove_outliers <- function(x, filltype) {
qnt <- quantile(x, probs=c(.25, .75), na.rm = TRUE)
 H <- 1.5 * IQR(x, na.rm = TRUE)
 y <- x
 #filltype=1 median
 #filltype=2 mean
 if (filltype ==1) {
y[x < (qnt[1] - H)] <- median(x,na.rm=TRUE)
 y[x > (qnt[2] + H)] <- median(x,na.rm=TRUE)
  }
else if (filltype ==2) {
   y[x < (qnt[1] - H)] <- mean(x,na.rm=TRUE)
 y[x > (qnt[2] + H)] <- mean(x,na.rm=TRUE)
 }
 else {
 y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
 }
  y
}
write.table(exposure, file = "Ndarexposure.txt", sep = ",", row.names = FALSE,col.names = TRUE )
```

As a result, after the preprocessing analysis, the “Exposure” table has 54 variables where 52 of them are continuous variables and 2 of them are categorical: Autistic and gender. The relation between the response variable, Autistic, and the other 53 independent variables can be shown below.

```
Autistic ~ fcc1_distance, ......, pm25_Trim1, pm25_Trim2, gender
```

Autism is a response and Bernoulli random variable which indicates the presence of Autism taking Y/N values. It is also binomially distributed.

6. IMPLEMENTATION & RESULTS

6.1 Implementation of Logistic Regression

6.1.1 Fitting a Logistic Regression Model

Whenever we wish to relate a collection independent variable to a dependent variable which is binary, we use multiple logistic regression or logit models. Firstly, we begin with complex multiple logistic regression model containing 53 independent variables as in (4.13)

$$logit(\pi) = ln\left[\frac{\pi}{1-\pi}\right] = \beta X' \qquad (6.1)$$

where $\pi = [1\ + exp(-\beta X)]^{-1}$ represents the probability of success of an event, $\beta = (\beta_0, \beta_1, \dots . \beta_{53})$ are the regression coefficients and X'denotes the independent variables (*fcc1_distance*, *fcc2_distance*, ... *roadtype4_Trim3_nox*, ... *pm25_Trim3 +gender)* in Table 3.10.

The model parameters are estimated or fitted to the data by the maximum likelihood (ML) method available in software packages such as RStudio, Minitab, and Stata. Thus, the resulting estimated parameters model fits the observed data most closely.

The fitting of the logistic regression model can be constructed in RStudio by using the g*lm()* function. The *glm()* function can be called with the arguments as in the following command.

```
> glm(formula:response ~ explanantory_variables,
               family= familyname(link= linkfunction)
```

In the *glm()* function, the family argument specifies the distributions and link functions to be used in the model. Family names ("binomial", "gaussian","Gamma", "poisson") can be used for logistic regression, linear regression, Gamma regression and Poisson regression respectively. Additionally, The most commonly link funtions can be (link = "logit"), "identity"), (link = "inverse") and (link = "log") for logistic regression, linear regression, gamma regression and poisson regression respectively. For example, *glm()* function for binomial distribution is

```
>glm(response ~ explanantory_variables, family=binomial(link="logit"),
data=mydata)
```

For an example, “Exposure” data was fitted by using *glm()* function and results were assigned to *model.full* object shown below:

Fitting of model.full

```
>model.full <- glm(Autistic  ~  fcc1_distance + fcc2_distance + fcc3_distance + fcc4_distance +
roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox +
roadtype1_Trim2_nox + roadtype1_Trim3_nox +  roadtype3_1stYr_nox + roadtype3_2ndYr_nox +
roadtype3_Preg_nox + roadtype3_Trim1_nox + roadtype3_Trim2_nox + roadtype3_Trim3_nox +
roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtype4_Preg_nox + roadtype4_Trim1_nox +
roadtype4_Trim2_nox + roadtype4_Trim3_nox +  roadtypeAll_1stYr_nox + roadtypeAll_2ndYr_nox
+ roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + roadtypeAll_Trim2_nox +
roadtypeAll_Trim3_nox +  no2_1stYr + no2_2ndYr + no2_Preg + no2_Trim1 + no2_Trim2 +
no2_Trim3 + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + o3_Trim2 + o3_Trim3 + pm10_1stYr +
pm10_2ndYr +  pm10_Preg + pm10_Trim1 + pm10_Trim2 + pm10_Trim3 +  pm25_1stYr +
pm25_2ndYr + pm25_Preg + pm25_Trim1 + pm25_Trim2 + pm25_Trim3 +  gender,
data=Exposure,family=binomial())
```

Then, the details of the *model.full* were obtained by using the *summary(model.full) R* command and the results were shown in Table 6.1.

```
> summary(model.full)
```

Table 6.1: Summary results of the fitting *model.full*

<table>
<tr><td></td><td colspan="6">Call:</td></tr>
<tr><td>Regression Equation</td><td colspan="6"><i>glm(formula = Autistic ~ fcc1_distance + fcc2_distance + fcc3_distance + fcc4_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_2ndYr_nox + roadtype3_Preg_nox + roadtype3_Trim1_nox + roadtype3_Trim2_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtype4_Preg_nox + roadtype4_Trim1_nox + roadtype4_Trim2_nox + roadtype4_Trim3_nox + roadtypeAll_1stYr_nox + roadtypeAll_2ndYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + roadtypeAll_Trim2_nox + roadtypeAll_Trim3_nox + no2_1stYr + no2_2ndYr + no2_Preg + no2_Trim1 + no2_Trim2 + no2_Trim3 + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + o3_Trim2 + o3_Trim3 + pm10_1stYr + pm10_2ndYr + pm10_Preg + pm10_Trim1 + pm10_Trim2 + pm10_Trim3 + pm25_1stYr + pm25_2ndYr + pm25_Preg + pm25_Trim1 + pm25_Trim2 + pm25_Trim3 + gender, family = binomial(), data = exposure)</i></td></tr>
<tr><td></td><td colspan="6"><i>Deviance Residuals:
Min 1Q Median 3Q Max
-2.4846 -1.1310 0.6377 0.9530 1.7378</i></td></tr>
<tr><td rowspan="19">Coefficients</td><td><b>Variables</b></td><td><b>Est. Coef.</b></td><td><b>Std.Error</b></td><td><b>z value</b></td><td><b>Pr(>|z|)</b></td><td><b>Signif. Codes</b></td></tr>
<tr><td>(Intercept)</td><td>-0.886200</td><td>0.844900</td><td>-1.049000</td><td>0.294262</td><td></td></tr>
<tr><td><b>fcc1_distance</b></td><td><b>-0.000060</b></td><td><b>0.000067</b></td><td><b>-0.906000</b></td><td><b>0.364998</b></td><td></td></tr>
<tr><td>fcc2_distance</td><td>-0.000020</td><td>0.000009</td><td>-2.181000</td><td>0.029188</td><td>*</td></tr>
<tr><td>fcc3_distance</td><td>0.000091</td><td>0.000450</td><td>0.203000</td><td>0.839084</td><td></td></tr>
<tr><td>fcc4_distance</td><td>0.006950</td><td>0.009590</td><td>0.725000</td><td>0.468586</td><td></td></tr>
<tr><td><b>roadtype1_1stYr_nox</b></td><td><b>0.191000</b></td><td><b>0.060490</b></td><td><b>3.158000</b></td><td><b>0.001588</b></td><td>**</td></tr>
<tr><td>roadtype1_2ndYr_nox</td><td>-0.157500</td><td>0.055020</td><td>-2.863000</td><td>0.004202</td><td>**</td></tr>
<tr><td>roadtype1_Preg_nox</td><td>-0.107100</td><td>0.040090</td><td>-2.670000</td><td>0.007575</td><td>**</td></tr>
<tr><td>roadtype1_Trim1_nox</td><td>-0.087450</td><td>0.025520</td><td>-3.426000</td><td>0.000612</td><td>***</td></tr>
<tr><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td></td></tr>
<tr><td>roadtype4_1stYr_nox</td><td>0.305500</td><td>0.157600</td><td>1.939000</td><td>0.052552</td><td>.</td></tr>
<tr><td>roadtype4_2ndYr_nox</td><td>-0.248100</td><td>0.147800</td><td>-1.678000</td><td>0.093260</td><td>.</td></tr>
<tr><td><b>roadtype4_Preg_nox</b></td><td><b>0.098540</b></td><td><b>0.119700</b></td><td><b>0.823000</b></td><td><b>0.410294</b></td><td></td></tr>
<tr><td>roadtype4_Trim1_nox</td><td>-0.035210</td><td>0.070750</td><td>-0.498000</td><td>0.618704</td><td></td></tr>
<tr><td>.</td><td>.</td><td>.</td><td>.</td><td>.</td><td></td></tr>
<tr><td><b>o3_Trim1</b></td><td><b>0.054880</b></td><td><b>0.024030</b></td><td><b>2.284000</b></td><td><b>0.022371</b></td><td>*</td></tr>
<tr><td>pm25_Preg</td><td>0.055890</td><td>0.062150</td><td>0.899000</td><td>0.368513</td><td></td></tr>
<tr><td><b>pm25_Trim1</b></td><td><b>0.024790</b></td><td><b>0.030140</b></td><td><b>0.822000</b></td><td><b>0.410851</b></td><td></td></tr>
<tr><td></td><td><b>genderMALE</b></td><td><b>0.351100</b></td><td><b>0.228500</b></td><td><b>1.537000</b></td><td><b>0.124321</b></td><td></td></tr>
<tr><td></td><td colspan="6">Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 983.31 on 750 degrees of freedom Residual deviance: 875.53 on 697 degrees of freedom AIC: 983.53</td></tr>
</table>

From the Table 6.1, the estimated coefficients for the *intercept*, *fcc1_distance*, *roadtype1_1stYr_nox*, *o3_Trim1*, *pm25_Trim1* and *genderMALE* variables' coefficients were -0.88620, -0.00006, 0.191, 0.05488, 0.02479 and 0.3511 respectively. For *the* *pm25_Trim1* variable, by using Eq. 4.36, it was interpreted that one-unit change in pm25_Trim1 resulted in a 0.02479 increase in log-odds.

$$log(OR_{pm25\_Trim1}) = 0.02479 \tag{6.2}$$

Thus, the estimated logit model containing 53 independent variables was written shortly as

$$Logit(\pi) = -0.8862 - 0.00006 * fcc1_{distance} + \cdots + 0.35 * genderMALE \tag{6.3}$$

6.1.2 Testing For the Significance of Coefficients with Wald Test

After fitting the model, estimating the coefficients, the Wald test can be used to test the significance of individual variables in the model. By using the values shown in the second and third columns in Table 6.1, labeled Estimate and Std.Error, in Eq. 4.46, Walt test for the coefficient of *roadtype1_Trim1_nox* variable was calculated as

$$z_{roadtype1\_Trim1\_nox} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{0.191}{0.06049} = 3.158\ and\ P(|z| > 3.158\ ) = 0.001588 \tag{6.4}$$

Walt test results, labeled *z*, for the coefficients were provided in the fourth column of Table 6.1. Also, the *p-value* in the fifth column, P(|z|>3.158) =0.001588 for |z|=3.158, was used to test the following null hypothesis H_o for the coefficient of *roadtype1_Trim1_nox*

$$H_0: \beta_{roadtype1\_Trim1\_nox} = 0\ versus\ \ H_1: \beta_{roadtype1\_Trim1\_nox} \neq 0 \tag{6.5}$$

Note that *p-values* are the smallest level of significance that leads to rejecting the null hypothesis H_0. Therefore, H_0 was rejected and the coefficient $\beta_{roadtype1\_Trim1\_nox}$ was significant since $\alpha > p - value = 0.000612$ for a level of significance ∝= 0.05.

As a result , for the *p-values* in fifth column in Table 6.1, the following conclusion was reached that *fcc2_distance, roadtype1_1stYr_nox, roadtype1_2ndYr_nox, o3_1stYr roadtype1_Preg_nox, roadtype1_Trim1_nox, roadtype3_Trim3_nox , o3_Trim1, roadtypeAll_Trim1_nox ,o3_2ndYr, and roadtype4_1stYr_nox variables* with (p<=0.05) values made a significant contribution to regression. Additionally, *roadtype1_Trim2_nox, roadtype3_1stYr_nox, o3_Preg , roadtype4_2ndYr_nox and roadtype3_Preg_nox variables* with *p-values* (0.05<p<=0.1) values made small contribution to regression. On the other hand, other variables made no contribution or very little contribution to regresssion

6.1.3 Interpreting the Model Coefficients

As we remember from chapter (4.2.6), for multivariate model with p variable that β_j coefficient of any focused variable X_j , represents the change in the log odds per unit change in a single factor X_j when all other factors are held constant as in Eq. 4.38, that is, $log(odds_2) - log(odds_1) = log(odds_2/odds_1) = log(OR) = \beta_j$. Recall from Eq. 4.40, a relationship between the odds ratio and the jth regression coefficient, β_j, is obtained. The odds ratio which is the relative increase in the odds of when X_j increases from *m* to *m+1* holding other variables fixed is

$$OR = \frac{Odds\ when\ X_j = m+1}{Odds\ when\ to\ X_j = m} = e^{\beta_j} \tag{6.6}$$

As an example, using estimated coefficients in Table 6.1 and Eq. 4.40, the odds ratios for *roadtype1_1stYr_nox* and *roadtype4_Preg_nox* variables were

$$OR_{roadtype1\_1stYr\_nox} = \exp(0.191) = 1.2105 \text{ , and}$$

$$OR_{roadtype4\_Preg\_nox} = \exp(0.09854) = 1.1036.$$

The OR=1.2105 value for *roadtype1_1stYr_nox* was interpreted that the odds of autism risk increased by a factor of exp(0.191)= 1.2105 for every increase of one unit in *roadtype1_1stYr_nox.* Since OR=1.2105 > 1 then it was also interpreted as having autism risk. If the *jth* coefficient β_j were 0, then the odds ratio, $OR = e^0$ would equal to 1 (no risk of autism).

The odds ratios were obtained by using the following R command:

```
> exp(coef(model.full))
```

Additionally, the 95% confidence interval for the odds ratio, $exp\ (\beta_{roadtype1\_Trim1\_nox})$, obtained from Eq. 4.48 was

$$[exp(0.191 - 1.96 * 0.06049), exp(0.191 + 1.96 * 0.06049)] = [1.0751, 1.3628]$$

Therefore

$$1.0751 \le OR \ll 1.3628$$

Since the interval did not contain one, the odds ratio for *roadtype1_1stYr_nox variable* was considered statistically significant at the 0.05 level. The above "hand calculation" was similar to confidence intervals provided by R software in Table 6.2. If desired, confidence intervals could be provided by R for the coefficients by using the *confint()* function. Confidence intervals for the odds ratio were also easily obtained with the following command such as odds ratio and percent confidence intervals for each of the coefficients as shown in Table 6.2.

```
> exp(cbind(OR= coef(model.full),confint(model.full)))
```

Table 6.2: Odds ratios(OR) and confidence intervals(CI) results for each coefficient of the *model.full*

Variables	OR	2.50%CI	97.50%CI
(Intercept)	0.413342	0.078313	2.161130
fcc1_distance	0.999939	0.999808	1.000070
fcc2_distance	0.999980	0.999962	0.999998
fcc3_distance	1.000092	0.999214	1.000980
fcc4_distance	1.006999	0.988283	1.026223
roadtype1_1stYr_nox	1.210854	1.081216	1.373152
roadtype1_2ndYr_nox	0.854000	0.761254	0.947551
roadtype1_Preg_nox	0.898151	0.827843	0.969262
.	.	.	.
.	.	.	.
o3_2ndYr	1.075196	1.009243	1.147661
o3_Preg	0.905062	0.793909	0.998888
o3_Trim1	1.057785	1.012896	1.114706
o3_Trim2	1.019428	0.979208	1.071390
o3_Trim3	1.024329	0.984914	1.072511
pm25_Trim1	1.025165	0.966328	1.087960
pm25_Trim2	0.960504	0.896831	1.027977
pm25_Trim3	0.979095	0.920300	1.043290
genderMALE	1.420802	0.906319	2.223800

6.1.4 Likelihood Ratio Test (LRT): Test Whether Several $\beta_k = 0$

The LRT can be used to examine whether a significant relationship exists between the dependent variable and the independent variable(s) contained in the logistic model. Therefore, the LRT was used to test the hypothesis that a few independent variables were zero or not. As an example, by dropping the variables whose p-values > 0.15 from the *model.full* above mentioned and a reduced model, *model.reduced,* was obtained and fitted with the following command.

Fitting of model.reduced

```
> model.reduced <- glm(formula = Autistic ~ fcc2_distance + roadtype1_1stYr_nox +
roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox
+ roadtype1_Trim3_nox + roadtype4_2ndYr_nox + roadtype3_1stYr_nox +
roadtype3_Preg_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtypeAll_Preg_nox
+ roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 +
pm25_2ndYr + pm25_Trim2 + gender, family = binomial(link = "logit"), data = Exposure)
```

Then, *model.reduce* and *model.ful* were compared manually with the following commands by using the Eq. 4.44 for the test static, χ^2, where L_R is the likelihood of *model.reduce* and L_F is the likelihood of *model.full*.

Likelihood Ratio Test

```
> -2* logLik(model.full)
'log Lik.' 875.1574 (df=54)

> -2* logLik(model.reduced)
'log Lik.' 913.3667 (df=17)

> x2= -2*logLik(model.reduced)- (-2*logLik(model.full)) # log-likelihood ratio test statistic

> as.numeric(x2)
[1] 38.20924

> as.numeric(pval=1-pchisq(x2,37))
[1] 0.4143158
```

By using *lmtest* package in R, a similar calculation with LRT was performed with the following code as

Likelihood Ratio Test

```
> library(lmtest)

> lrtest(model.reduced,model.full) # Likelihood Ratio Test
Model 1: Autistic ~ fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox +
roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype3_Trim3_nox + roadtypeAll_Trim1_nox +
o3_2ndYr + o3_Trim1 + roadtype1_Trim2_nox + roadtype3_Preg_nox + roadtype3_1stYr_nox +
roadtype1_Trim2_nox + roadtype3_1stYr_nox + roadtype4_1stYr_nox + roadtype4_2ndYr_nox +
o3_1stYr + o3_Preg
Model 2: Autistic ~ fcc1_distance + fcc2_distance + fcc3_distance + fcc4_distance +
roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox +
roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_2ndYr_nox +
roadtype3_Preg_nox +  roadtype3_Trim1_nox + roadtype3_Trim2_nox + roadtype3_Trim3_nox +
roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtype4_Preg_nox + roadtype4_Trim1_nox +
roadtype4_Trim2_nox + roadtype4_Trim3_nox + roadtypeAll_1stYr_nox + roadtypeAll_2ndYr_nox
+ roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + roadtypeAll_Trim2_nox +
roadtypeAll_Trim3_nox + no2_1stYr + no2_2ndYr + no2_Preg + no2_Trim1 + no2_Trim2 +
no2_Trim3 + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + o3_Trim2 + o3_Trim3 + pm10_1stYr +
pm10_2ndYr + pm10_Preg + pm10_Trim1 + pm10_Trim2 + pm10_Trim3 + pm25_1stYr +
pm25_2ndYr + pm25_Preg + pm25_Trim1 + pm25_Trim2 + pm25_Trim3 + gender

  #Df  LogLik Df  Chisq Pr(>Chisq)
1  17 -456.68
2  54 -437.58 37 38.209     0.4143
```

As we see from the result of LRT, the full model contained 54 variables, the reduced model contained 17 variables and the *p-value* for a chi-square value of 38.21 with 37 degrees of freedom was 0.4143. When the computed $\chi^2 = 38.21$ value was compared with the percentage quartile point $\chi^2{}_{(1-0.05,37)} = 52.19$, from a chi-square distribution table, it was seen that $\chi^2 \leq \chi^2{}_{(1-0.05,37)}$ or $P(\geq \chi^2) = 0.4143$, which was insignificant at the α = 0.05 level. According to Eq. 4.45a we accepted the null hypothesis, H_0, and some coefficients of the variables were zero and insignificant. Therefore, we concluded that the reduced model was better than the full model.

6.2 Variable Selection

6.2.1 Purposeful Variables Selection

The purposeful variable selection algorithm explained in Chapter (4.2.8.4) was performed by the following steps.

Step-1*:* The full model, `model.full,` that contains all variables, was fitted with the “Exposure” data by using the *glm()* function with the following command.

<table>
<tr><td><b>Fitting of model.full</b></td><td><i>&gt;model.full = glm(Autistic ~ fcc1_distance + fcc2_distance + fcc3_distance + fcc4_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + o3_Trim3 + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_2ndYr_nox + roadtype3_Preg_nox + o3_Trim2 + roadtype3_Trim1_nox + roadtype3_Trim2_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + o3_Trim1 + roadtype4_2ndYr_nox + roadtype4_Preg_nox + roadtype4_Trim1_nox + roadtype4_Trim2_nox + o3_Preg + roadtype4_Trim3_nox + roadtypeAll_1stYr_nox + roadtypeAll_2ndYr_nox + roadtypeAll_Preg_nox + o3_2ndYr + roadtypeAll_Trim1_nox + roadtypeAll_Trim2_nox + roadtypeAll_Trim3_nox + no2_1stYr + no2_2ndYr + no2_Preg + no2_Trim1 + no2_Trim2 + no2_Trim3 + o3_1stYr + pm10_1stYr + pm10_2ndYr + pm10_Preg + pm10_Trim1 + pm10_Trim2 + pm10_Trim3 + pm25_1stYr + pm25_2ndYr + pm25_Preg + pm25_Trim1 + pm25_Trim2 + pm25_Trim3 + gender, data=Exposure, family=binomial())</i></td></tr>
</table>

Then, the results of this analysis were obtained with the *summary(model.full)* command and shown in Table 6.3.

```
> summary(model.full)
```

Table 6.3: Results of fitting of the *model.full*

<table>
<tr><td></td><td colspan="6"><code>Call:</code></td></tr>
<tr><td><b>Regression Equation</b></td><td colspan="6"><i>glm(formula = Autistic ~ fcc1_distance + fcc2_distance + fcc3_distance + fcc4_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_2ndYr_nox + roadtype3_Preg_nox + roadtype3_Trim1_nox + roadtype3_Trim2_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtype4_Preg_nox + roadtype4_Trim1_nox + roadtype4_Trim2_nox + roadtype4_Trim3_nox + roadtypeAll_1stYr_nox + roadtypeAll_2ndYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + roadtypeAll_Trim2_nox + roadtypeAll_Trim3_nox + no2_1stYr + no2_2ndYr + no2_Preg + no2_Trim1 + no2_Trim2 + no2_Trim3 + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + o3_Trim2 + o3_Trim3 + pm10_1stYr + pm10_2ndYr + pm10_Preg + pm10_Trim1 + pm10_Trim2 + pm10_Trim3 + pm25_1stYr + pm25_2ndYr + pm25_Preg + pm25_Trim1 + pm25_Trim2 + pm25_Trim3 + gender, family = binomial(), data = exposure)</i></td></tr>
<tr><td></td><td colspan="6"><i>Deviance Residuals:
Min   1Q   Median   3Q   Max
-2.4846 -1.1310   0.6377   0.9530   1.7378</i></td></tr>
<tr><td rowspan="11"><b>Coefficients</b></td><td><b>Variables</b></td><td><b>Estimate</b></td><td><b>Std.Error</b></td><td><b>z value</b></td><td><b>Pr(&gt;|z|)</b></td><td><b>Signif. Codes</b></td></tr>
<tr><td>(Intercept)</td><td>-0.8862</td><td>0.8449</td><td>-1.0490</td><td>0.2943</td><td></td></tr>
<tr><td>fcc1_distance</td><td>-0.0001</td><td>0.0001</td><td>-0.9060</td><td>0.3650</td><td></td></tr>
<tr><td>fcc2_distance</td><td>0.0000</td><td>0.0000</td><td>-2.1810</td><td>0.0292</td><td>*</td></tr>
<tr><td>fcc3_distance</td><td>0.0001</td><td>0.0004</td><td>0.2030</td><td>0.8391</td><td></td></tr>
<tr><td>fcc4_distance</td><td>0.0070</td><td>0.0096</td><td>0.7250</td><td>0.4686</td><td></td></tr>
<tr><td>roadtype1_1stYr_nox</td><td>0.1910</td><td>0.0605</td><td>3.1580</td><td>0.0016</td><td>**</td></tr>
<tr><td>roadtype1_2ndYr_nox</td><td>-0.1575</td><td>0.0550</td><td>-2.8630</td><td>0.0042</td><td>**</td></tr>
<tr><td>roadtype1_Preg_nox</td><td>-0.1071</td><td>0.0401</td><td>-2.6700</td><td>0.0076</td><td>**</td></tr>
<tr><td>roadtype1_Trim1_nox</td><td>-0.0875</td><td>0.0255</td><td>-3.4260</td><td>0.0006</td><td>***</td></tr>
<tr><td>roadtype1_Trim2_nox</td><td>0.0528</td><td>0.0298</td><td>1.7710</td><td>0.0765</td><td>.</td></tr>
</table>

	roadtype1_Trim3_nox	0.0489	0.0319	1.5330	0.1253	
	roadtype3_1stYr_nox	0.1250	0.0712	1.7560	0.0790	.
	roadtype3_2ndYr_nox	-0.0340	0.0648	-0.5250	0.5998	
	roadtype3_Preg_nox	-0.0874	0.0534	-1.6370	0.1017	
	roadtype3_Trim1_nox	-0.0049	0.0355	-0.1370	0.8907	
	roadtype3_Trim2_nox	0.0200	0.0451	0.4440	0.6571	
	roadtype3_Trim3_nox	-0.0888	0.0414	-2.1450	0.0320	*
	roadtype4_1stYr_nox	0.3055	0.1576	1.9390	0.0526	.
	roadtype4_2ndYr_nox	-0.2481	0.1478	-1.6780	0.0933	.
	roadtype4_Preg_nox	0.0985	0.1197	0.8230	0.4103	
	roadtype4_Trim1_nox	-0.0352	0.0708	-0.4980	0.6187	
	roadtype4_Trim2_nox	-0.0837	0.0916	-0.9130	0.3610	
	roadtype4_Trim3_nox	0.0255	0.0788	0.3240	0.7459	
	roadtypeAll_1stYr_nox	-0.0061	0.0265	-0.2300	0.8178	
	roadtypeAll_2ndYr_nox	0.0122	0.0237	0.5130	0.6076	
	roadtypeAll_Preg_nox	0.0272	0.0203	1.3400	0.1803	
	roadtypeAll_Trim1_nox	0.0463	0.0176	2.6270	0.0086	**
	roadtypeAll_Trim2_nox	-0.0150	0.0158	-0.9520	0.3412	
	roadtypeAll_Trim3_nox	0.0006	0.0174	0.0370	0.9707	
	no2_1stYr	0.0521	0.0769	0.6770	0.4983	
	no2_2ndYr	-0.0896	0.0653	-1.3710	0.1703	
	no2_Preg	-0.0114	0.0720	-0.1590	0.8739	
	no2_Trim1	0.0097	0.0346	0.2800	0.7794	
	no2_Trim2	-0.0255	0.0332	-0.7680	0.4428	
	no2_Trim3	-0.0044	0.0332	-0.1310	0.8955	
	o3_1stYr	-0.0733	0.0374	-1.9590	0.0501	.
	o3_2ndYr	0.0729	0.0327	2.2280	0.0259	*
	o3_Preg	-0.0971	0.0576	-1.6870	0.0916	.
	o3_Trim1	0.0549	0.0240	2.2840	0.0224	*
	o3_Trim2	0.0185	0.0225	0.8210	0.4116	
	o3_Trim3	0.0229	0.0214	1.0720	0.2837	
	pm10_1stYr	0.0120	0.0341	0.3520	0.7247	
	pm10_2ndYr	0.0169	0.0344	0.4910	0.6236	
	pm10_Preg	-0.0409	0.0444	-0.9190	0.3578	
	pm10_Trim1	-0.0257	0.0245	-1.0520	0.2929	
	pm10_Trim2	0.0287	0.0256	1.1230	0.2613	
	pm10_Trim3	0.0076	0.0220	0.3450	0.7301	
	pm25_1stYr	0.0489	0.0541	0.9040	0.3660	
	pm25_2ndYr	0.0775	0.0606	1.2780	0.2012	
	pm25_Preg	0.0559	0.0622	0.8990	0.3685	
	pm25_Trim1	0.0248	0.0301	0.8220	0.4109	
	pm25_Trim2	-0.0392	0.0346	-1.1330	0.2573	
	pm25_Trim3	-0.0210	0.0319	-0.6580	0.5107	
	genderMALE	0.3511	0.2285	1.5370	0.1243	
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 983.31 on 750 degrees of freedom Residual deviance: 875.53 on 697 degrees of freedom AIC: 983.53					

The selected variables whose p-values < 0.25, moderately associated with the response, were shown in Table 6.4 for the next step.

Table 6.4: The list of selected variables that are significant at the 0.25 level.

Variables	Est. Coef.	Std.Error	z value	Pr(>|z|)	Signif. Codes
fcc2_distance	0.0000	0.0000	-2.1810	0.0292	*
roadtype1_1stYr_nox	0.1910	0.0605	3.1580	0.0016	**
roadtype1_2ndYr_nox	-0.1575	0.0550	-2.8630	0.0042	**
roadtype1_Preg_nox	-0.1071	0.0401	-2.6700	0.0076	**
roadtype1_Trim1_nox	-0.0875	0.0255	-3.4260	0.0006	***
roadtype1_Trim2_nox	0.0528	0.0298	1.7710	0.0765	.
roadtype1_Trim3_nox	0.0489	0.0319	1.5330	0.1253	
roadtype3_1stYr_nox	0.1250	0.0712	1.7560	0.0790	.
roadtype3_Preg_nox	-0.0874	0.0534	-1.6370	0.1017	
roadtype3_Trim3_nox	-0.0888	0.0414	-2.1450	0.0320	*
roadtype4_1stYr_nox	0.3055	0.1576	1.9390	0.0526	.
roadtype4_2ndYr_nox	-0.2481	0.1478	-1.6780	0.0933	.
roadtypeAll_Preg_nox	0.0272	0.0203	1.3400	0.1803	
roadtypeAll_Trim1_nox	0.0463	0.0176	2.6270	0.0086	**
no2_2ndYr	-0.0896	0.0653	-1.3710	0.1703	
o3_1stYr	-0.0733	0.0374	-1.9590	0.0501	.
o3_2ndYr	0.0729	0.0327	2.2280	0.0259	*
o3_Preg	-0.0971	0.0576	-1.6870	0.0916	.
o3_Trim1	0.0549	0.0240	2.2840	0.0224	*
pm25_2ndYr	0.0775	0.0606	1.2780	0.2012	
pm25_Trim2	-0.0392	0.0346	-1.1330	0.2573	
genderMALE	0.3511	0.2285	1.5370	0.1243	

***Step-2**:* Our first reduced model, *model.1*, was formed with the selected variables in Table 6.4 and fitted with the following command.

Fitting of model.full

```
> model.1 <- glm(formula = Autistic ~ fcc2_distance + roadtype1_1stYr_nox +
roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox +
roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype4_2ndYr_nox +
roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox +
roadtype4_1stYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr
+ o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + pm25_2ndYr + pm25_Trim2 + gender,
family = binomial(link = "logit"), data = Exposure)
```

Then, the details of the fitted model, *model.1*, were obtained with the *summary(model.1)* command, and the results of it are shown in Table 6.5.

```
> summary(model.1)
```

Table 6.5: Results of fitting of *model.1*

	`Call:`					
Regression Equation	*glm(formula = Autistic ~ fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype4_2ndYr_nox + roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + pm25_2ndYr + pm25_Trim2 + gender, family = binomial(link = "logit"), data = Exposure)*					
	Deviance Residuals: *Min 1Q Median 3Q Max* *-2.4037 -1.1548 0.6563 0.9637 1.5995*					
Coefficients	**Variables**	**Est. Coef.**	**Std.Error**	**z value**	**Pr(>|z|)**	**Signif. Codes**
	(Intercept)	-0.7806	0.7409	-1.0540	0.2921	
	fcc2_distance	0.0000	0.0000	-2.3370	0.0194	*
	roadtype1_1stYr_nox	0.1809	0.0525	3.4430	0.0006	***
	roadtype1_2ndYr_nox	-0.1423	0.0493	-2.8850	0.0039	**
	roadtype1_Preg_nox	-0.1037	0.0353	-2.9360	0.0033	**
	roadtype1_Trim1_nox	-0.0705	0.0234	-3.0080	0.0026	**
	roadtype1_Trim2_nox	0.0397	0.0242	1.6370	0.1015	
	roadtype1_Trim3_nox	0.0556	0.0265	2.0980	0.0359	*
	roadtype4_2ndYr_nox	-0.2139	0.1368	-1.5640	0.1178	
	roadtype3_1stYr_nox	0.1036	0.0396	2.6160	0.0089	**
	roadtype3_Preg_nox	-0.0740	0.0374	-1.9770	0.0480	*
	roadtype3_Trim3_nox	-0.0977	0.0343	-2.8470	0.0044	**
	roadtype4_1stYr_nox	0.2740	0.1353	2.0260	0.0428	*
	roadtypeAll_Preg_nox	0.0226	0.0140	1.6170	0.1058	
	roadtypeAll_Trim1_nox	0.0399	0.0145	2.7630	0.0057	**
	no2_2ndYr	-0.0669	0.0355	-1.8820	0.0599	.
	o3_1stYr	-0.0605	0.0325	-1.8610	0.0627	.
	o3_2ndYr	0.0662	0.0296	2.2340	0.0255	*
	o3_Preg	-0.0238	0.0190	-1.2550	0.2096	
	o3_Trim1	0.0208	0.0098	2.1300	0.0332	*
	pm25_2ndYr	0.1462	0.0430	3.4010	0.0007	***
	pm25_Trim2	-0.0252	0.0198	-1.2740	0.2026	
	genderMALE	0.3645	0.2198	1.6580	0.0973	.
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 983.31 on 750 degrees of freedom Residual deviance: 891.98 on 728 degrees of freedom AIC: 937.98					

The significance of each variable was examined by using the p-value of the Walt test shown in the fifth column of Table 6.5. Then, any variable with the p-values greater than $\alpha_{cri.}$=0.15 was eliminated from *model.1*. At the end of the fitting of the *model.1*, *o3_Preg and pm25_Trim2* variables with (p>=0.15) were removed from the *model.1*.

Next, *model.2* was formed and fitted without *o3_Preg and pm25_Trim2* variables. Then, the details of the fitted *model.2* were shown in Table 6.6 by using the *summary(model.2)* command.

```
> summary(model.2) # display results
```

Table 6.6: Results of fitting of *model.2*

	Call:					
Regression Equation	*glm(formula = Autistic ~ fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype4_2ndYr_nox + roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr + o3_2ndYr + o3_Trim1 + pm25_2ndYr + gender, family = binomial(), data = Exposure)*					
	Deviance Residuals: *Min 1Q Median 3Q Max* *-2.4037 -1.1548 0.6563 0.9637 1.5995*					
Coefficients	**Variables**	**Est. Coef.**	**Std.Error**	**z value**	**Pr(>|z|)**	**Signif. Codes**
	(Intercept)	-0.9033	0.7349	-1.2290	0.2190	
	fcc2_distance	0.0000	0.0000	-2.2310	0.0257	*
	roadtype1_1stYr_nox	0.1738	0.0519	3.3510	0.0008	***
	roadtype1_2ndYr_nox	-0.1383	0.0487	-2.8420	0.0045	**
	roadtype1_Preg_nox	-0.1024	0.0353	-2.9060	0.0037	**
	roadtype1_Trim1_nox	-0.0664	0.0232	-2.8650	0.0042	**
	roadtype1_Trim2_nox	0.0373	0.0239	1.5600	0.1187	
	roadtype1_Trim3_nox	0.0568	0.0266	2.1310	0.0331	*
	roadtype4_2ndYr_nox	-0.2038	0.1362	-1.4960	0.1345	
	roadtype3_1stYr_nox	0.1005	0.0393	2.5610	0.0104	*
	roadtype3_Preg_nox	-0.0739	0.0368	-2.0070	0.0447	*
	roadtype3_Trim3_nox	-0.0913	0.0340	-2.6820	0.0073	**
	roadtype4_1stYr_nox	0.2573	0.1344	1.9150	0.0555	.
	roadtypeAll_Preg_nox	0.0234	0.0140	1.6720	0.0945	.
	roadtypeAll_Trim1_nox	0.0365	0.0143	2.5590	0.0105	*
	no2_2ndYr	-0.0669	0.0355	-1.8850	0.0594	.
	o3_1stYr	-0.0761	0.0296	-2.5740	0.0101	*
	o3_2ndYr	0.0670	0.0297	2.2600	0.0238	*
	o3_Trim1	0.0129	0.0081	1.5860	0.1128	
	pm25_2ndYr	0.1269	0.0398	3.1900	0.0014	**
	genderMALE	0.3794	0.2194	1.7290	0.0838	.
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 983.31 on 750 degrees of freedom Residual deviance: 894.17 on 730 degrees of freedom AIC: 936.17					

Then, to compare *model.1* and *model.2,* LRT was performed with the following code as:

Like9ihood ratio test

```
> lrtest(model.2,model.1) #Likelihood ratio test
Model 1: Autistic ~ fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox +
roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox +
roadtype4_2ndYr_nox + roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox +
roadtype4_1stYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr +
o3_2ndYr + o3_Trim1 + pm25_2ndYr + gender
Model 2: Autistic ~ fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox +
roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox +
roadtype4_2ndYr_nox + roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox +
roadtype4_1stYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr +
o3_2ndYr + o3_Preg + o3_Trim1 + pm25_2ndYr + pm25_Trim2 + gender
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  21 -447.08
2  23 -445.99  2 2.1829     0.3357
```

According to the above LRT result:

Since $\chi^2 = 2.18 \leq \chi^2_{(1-0.05,2)} = 5.99$, or $P(\geq \chi^2) = 0.3357$, which was insignificant at the α = 0.05 level, the hypothesis H_0 with some zero coefficients, was accepted according to Eq. 4.45a. Therefore, we concluded that the *model.2* was better than *model.1* and all the variables in *model.2* had p-values < 0.15, that is, were significant at the 15% level.

Step-3: By using Eq. 4.51, the percentage changes in each coefficient of the reduced and large model were compared with the following code.

Calculation % Changes in each Coefficient

```
#Calculation % Changes in each Coefficent
coefpos=c(19,22)
coef1=coef(model.1)[- coefpos] #drop o3_Preg and pm25_Trim2
coef2=coef(model.2) #extracts coefficients from model.2
deltaB=100*abs((coef2-coef1)/coef1)
for(i in 1:21){
   cat(names(delta.coef)[i],"\t",unname(delta.coef)[i],"\n")
}
```

The largest percent changes were obtained for *o3_1stYr* and *o3_Trim1* variables which increased by 25.69% and 38.17% respectively, exceeded criterion of 20%. Thus, one or more of the excluded variables might be important and the remained model might need adjustment. But, it was concluded not to add back *o3_Preg* and *pm25_Trim2* variables into *model.2* based on the above LRT results, since the *model.2* (without *o3_1stYr* and *o3_Trim1* variables) was better than *model.1*. As a result, at the end of the fitting of *model.2,* all the variables had a *p-value* of less than 0.15. Then, *model.2* was called as the *main effects model*, which contained the important variables.

Step-4: In this step, each continuous variable in *model.3* were checked for their linearity relation to the logit of the outcome. So, in the R-code below, the *scatter.smooth()* function was used to display the relationship between two variables. Additionally, to visualize the relationship better, the *abline()* function was used to add a straight regression line or a smoothed curve to the scatter plot.

Checking Linearity of Variables to the Logit of the Outcome

```
# Checking the linearity of variables to the logit of the outcome
p=fitted(model.3)
logodds=log(p/(1-p))
test=summary(model.3)$coefficients
row=rownames(test)
d=dim(test)-1
par(mfrow=c(2,2))
for(i in 2:d[1]){
     col=row[i]
     coldata=eval(parse(text=paste("Exposure$", col, sep = "")))
       t=which(posrow==col)
       if (length(t)>0){
         x1=paste("glm(Autistic  ~ ", col,", data=Exposure, family=binomial)")
         fit=eval(parse(text = x1))
         scatter.smooth(coldata, logodds,xlab =col, lpars = list(col = "blue", lwd = 3, lty = 1))
         abline(fit, col="red", lwd=3, lty = 1)
       }
}
```

According to the results of scattering, plots of the above code, linearly associated variables were shown in Figure 6.1. while nonlinearly related variables were shown in Figure 6.2. Nonlinearly related variables also showed partial linearity between some values. So it can be accepted that logit linearly increases or decreases as a function of the variables

Step-5: In this step, firstly, all combinations of interactions among continuous variables at 2 degrees were added to the model instead of adding one by one and they were checked with the following code. Possible pairs of variables in the model were automatically included by R-code as the arithmetic product of the pairs of main effect variables in the *model.int* such as `Autistic ~(.)^2`. Then, the interactions whose p-values < 0.001, highly associated with response were selected, and the details of the fitted *model.int* were shown in Table 6.7. Finally, when the interactions in Table 6.7 were included in the model and refitted, interactions were found to be not significant with *p-values* (0.172, 0.265). So the model just contained main effects, not interactions.

Finding Interactions	*>model.int <- glm(Autistic ~ (fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr + o3_2ndYr + o3_Trim1 + pm25_2ndYr)^2, data=Exposure,family=binomial(link="logit"))*

Table 6.7: Results of fitting of *model.int*, the interactions whose p-values < 0.001

Variables	Est. Coef.	Std.Error	z value	Pr(>|z|)	Signif. Codes
roadtype3_Preg_nox:o3_Trim1	0.0432	0.0114	3.7900	0.0002	***
roadtype3_Trim3_nox:roadtypeAll_Trim1_nox	-0.0583	0.0165	-3.5440	0.0004	***

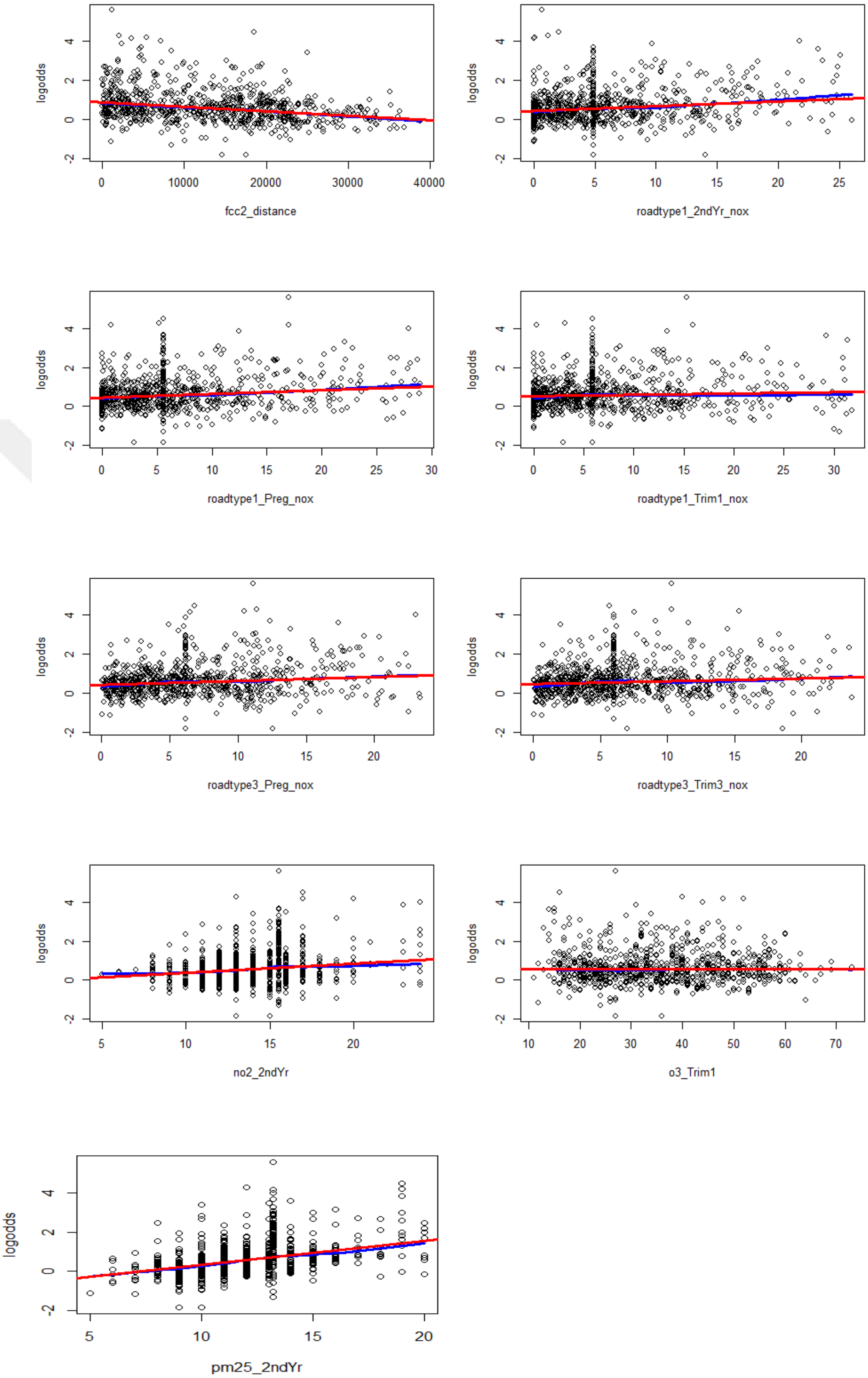


Figure 6.1: Scatter plots of linearly related variables with log-odds

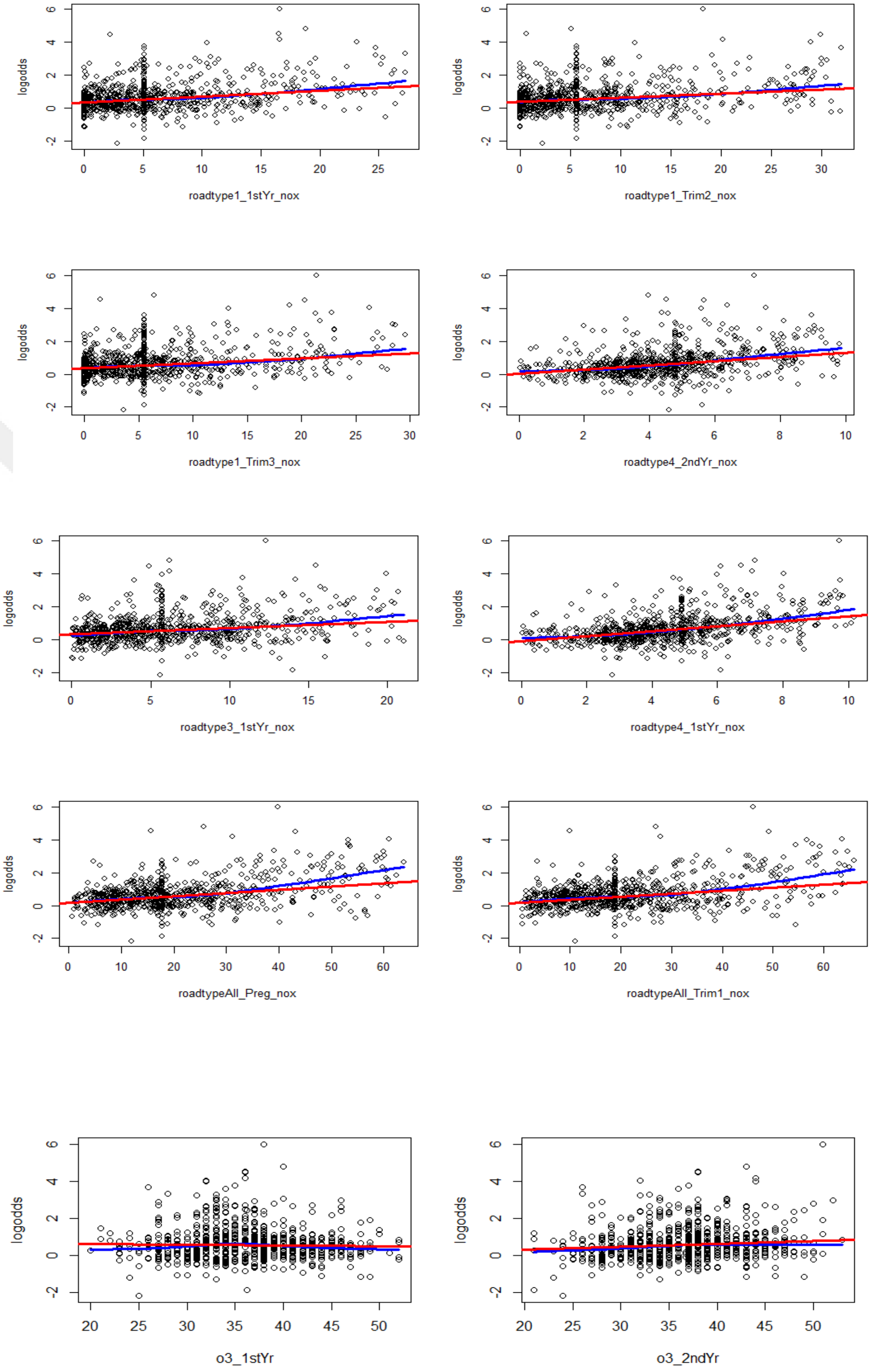


Figure 6.2: Scatter plots of nonlinearly related variables with log-odds

6.2.2 Stepwise Variable Selection

Stepwise Algorithms select the best subset of the predictors among many predictors for a model in a stepwise manner. Stepwise methods such as backward elimination, forward selection, and stepwise selection which combines both forward and backward can be used in model selection.

6.2.2.1 Stepwise Selection According to the P-value

Backward elimination is one of the stepwise selection algorithms. In this version, it begins with all variables in the model, and sequentially removes the variable that has the highest *p-value* greater than α_{crit}. It was implemented with the following code due to the *p-value* criterion. Then, the model was improved step by step by refitting and dropping a variable from the model until another step didn't show an improvement of the model fit [79, 89].

The α_{crit} is sometimes called the "*p-to-remove*" and does not have to be 5%. If prediction performance is the goal, then a 15 to 20% cutoff may work best" [97]. The results of the following *backward elimination* codes were shown in Table 6.8. As a result *backward elimination* dropped 33 variables for α_{crit}=0.15.

Backward Elimination According to P-value

```
#Backward elimination According to P-value
modelFormula="Autistic  ~  fcc1_distance + fcc2_distance + fcc3_distance + fcc4_distance +
roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox +
roadtype1_Trim2_nox + roadtype1_Trim3_nox +  roadtype3_1stYr_nox + roadtype3_2ndYr_nox +
roadtype3_Preg_nox + roadtype3_Trim1_nox + roadtype3_Trim2_nox + roadtype3_Trim3_nox +
roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtype4_Preg_nox + roadtype4_Trim1_nox +
roadtype4_Trim2_nox + roadtype4_Trim3_nox +  roadtypeAll_1stYr_nox +
roadtypeAll_2ndYr_nox + roadtypeAll_Preg_nox +   no2_1stYr + roadtypeAll_Trim1_nox +
roadtypeAll_Trim2_nox +  no2_2ndYr roadtypeAll_Trim3_nox + no2_Preg + no2_Trim1 +
no2_Trim2 + no2_Trim3 + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + o3_Trim2 + o3_Trim3 +
pm10_1stYr + pm10_2ndYr + pm10_Preg + pm10_Trim1 + pm10_Trim2 + pm10_Trim3 +
pm25_1stYr + pm25_2ndYr + pm25_Preg + pm25_Trim1 + pm25_Trim2 + pm25_Trim3 +  gender"
........
```

Backward Elimination According to P-value

```
.......
model.p <- glm(modelFormula, data=Exposure,family=binomial())
sig=0.15
counter=1
terms <- attr(model.p$terms,"term.labels")
while(T){
   test=summary(model.p)$coefficients
   pval <- test[,dim(test)[2]]
   names(pval) <- rownames(test)
   pval <- pval[names(pval)!="(Intercept)"]
   pval <- sort(pval,decreasing=T)
   if(sum(is.na(pval))>0) stop(paste("Model",deparse(substitute(model)),"is invalid. Check if all
coefficients are estimated."))
   if(pval[1]<sig)  # check if all significant
   {
     sprintf("Less than significant value ", counter)
     View(test)
     print(test)
     #break
    return()
     }
   dropvar <- names(pval)[1]
   if (dropvar=="genderMALE"){
     dropvar="gender"}
   terms <- terms[-match(dropvar,terms)]
   modelFormulas <- as.formula(paste(".~.-",dropvar))
   model.p <- update(model.p,modelFormulas)
   update(model.p)  #update terms and model
   #sprintf("model update = ", counter)
   cat("\n--------\nmodel updated",counter,"\n--------\n\n")
   if(length(terms)==0){
      sprintf("end of scopevars ", counter)
      View(test)
      print(test)
      return()
     }
   counter=counter+1
}
#Summary(model.p)
```

Table 6.8: The results of *Backward elimination* w.r.t. *p-value*

Variables	Est. Coef.	Std.Error	z value	Pr(>|z|)
fcc2_distance	0.0000	0.0000	-2.2305	0.0257
roadtype1_1stYr_nox	0.1738	0.0519	3.3505	0.0008
roadtype1_2ndYr_nox	-0.1383	0.0487	-2.8418	0.0045
roadtype1_Preg_nox	-0.1024	0.0353	-2.9058	0.0037
roadtype1_Trim1_nox	-0.0664	0.0232	-2.8654	0.0042
roadtype1_Trim2_nox	0.0373	0.0239	1.5603	0.1187
roadtype1_Trim3_nox	0.0567	0.0266	2.1313	0.0331
roadtype3_1stYr_nox	0.1005	0.0392	2.5612	0.0104
roadtype3_Preg_nox	-0.0739	0.0368	-2.0070	0.0447
roadtype3_Trim3_nox	-0.0913	0.0340	-2.6823	0.0073
roadtype4_1stYr_nox	0.2573	0.1344	1.9146	0.0555
roadtype4_2ndYr_nox	-0.2038	0.1362	-1.4964	0.1345
roadtypeAll_Preg_nox	0.0234	0.0140	1.6723	0.0945
roadtypeAll_Trim1_nox	0.0365	0.0143	2.5586	0.0105
no2_2ndYr	-0.0669	0.0355	-1.8854	0.0594
o3_1stYr	-0.0761	0.0296	-2.5740	0.0101
o3_2ndYr	0.0670	0.0297	2.2596	0.0238
o3_Trim1	0.0129	0.0081	1.5857	0.1128
pm25_2ndYr	0.1269	0.0398	3.1896	0.0014
genderMALE	0.3794	0.2194	1.7291	0.0838

6.2.2.2 Stepwise Selection According to Akai Information Criterion (AIC)

The *backward elimination* version of stepwise selection was performed by using *step()* function in R with direction=" backward" option to minimize the AIC value. It began with a complex model. At each step, AIC value was computed for each model that was formed by deleting a single variable from the current model and then the variable whose deletion results in the min AIC was removed from the model. The process stopped if another step didn't show a further improvement of the model. For my data, *model.s* was obtained with the *step()* function shown below.

Step() Function

```
> model.s= step(model.full,direction="backward", trace=1)
#trace=0 suppresses step by step output
```

The output of the *step()* function was shown in Table 6.9 with the trace=1 option.

Table 6.9: The output of the *step()* function with trace=1 option

Theoutput of Step() Function with Trace=1 Option

Start: AIC=983.53
Autistic ~ fcc1_distance + fcc2_distance + fcc3_distance + fcc4_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_2ndYr_nox + roadtype3_Preg_nox + roadtype3_Trim1_nox + roadtype3_Trim2_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtype4_Preg_nox + roadtype4_Trim1_nox + roadtype4_Trim2_nox + roadtype4_Trim3_nox + roadtypeAll_1stYr_nox + roadtypeAll_2ndYr_nox roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_1stYr + roadtypeAll_Trim2_nox + roadtypeAll_Trim3_nox + no2_2ndYr + no2_Preg + no2_Trim1 + no2_Trim2 + no2_Trim3 + o3_1stYr + o3_2ndYr + o3_Preg + o3_Trim1 + o3_Trim2 + o3_Trim3 + pm10_1stYr + pm10_2ndYr + pm10_Preg + pm10_Trim1 + pm10_Trim2 + pm10_Trim3 + pm25_1stYr + pm25_2ndYr + pm25_Preg + pm25_Trim1 + pm25_Trim2 + pm25_Trim3 + gender

	Variables	Df	Deviance	AIC
-	roadtypeAll_Trim3_nox	1	875.53	981.53
-	no2_Trim3	1	875.55	981.55
-	.	.	.	.
-	.	.	.	.
-	roadtype1_2ndYr_nox	1	884.58	990.58
-	roadtype1_1stYr_nox	1	887.12	993.12
-	roadtype1_Trim1_nox	1	888.85	994.85

.
< several more steps >
.
Final Step: AIC=936.18
Autistic ~ fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr + o3_2ndYr + o3_Trim1 + pm25_2ndYr + gender

	Variables	Df	Deviance	AIC
	<none>		894.18	936.18
-	roadtype4_2ndYr_nox	1	896.40	936.40
-	o3_Trim1	1	896.67	936.67
-	roadtype1_Trim2_nox	1	896.72	936.72
-	gender	1	897.15	937.15
-	roadtypeAll_Preg_nox	1	897.17	937.17
-	no2_2ndYr	1	897.73	937.73
-	roadtype4_1stYr_nox	1	897.82	937.82
-	roadtype3_Preg_nox	1	898.23	938.23
-	roadtype1_Trim3_nox	1	898.83	938.83
-	fcc2_distance	1	899.19	939.19
-	o3_2ndYr	1	899.46	939.46
-	o3_1stYr	1	900.99	940.99
-	roadtype3_1stYr_nox	1	901.14	941.14
-	roadtypeAll_Trim1_nox	1	901.28	941.28
-	roadtype3_Trim3_nox	1	901.48	941.48
-	roadtype1_Preg_nox	1	902.98	942.98
-	roadtype1_Trim1_nox	1	903.19	943.19
-	roadtype1_2ndYr_nox	1	903.34	943.34
-	pm25_2ndYr	1	904.72	944.72
-	roadtype1_1stYr_nox	1	907.70	947.70

According to the above backward elimination by using step() function, it started with the *model.full* and the initial AIC value is 983.53. At each step, one variable was deleted at a time then AIC value was computed. For example, when *roadtypeAll_Trim3_nox* and *no2_Trim3* variables were deleted sequentially, 981.53, and 981.55 AIC values were computed respectively. Since the AIC value of *roadtypeAll_Trim3_nox* was the smallest, it was dropped from the model and the next step began. Then the process repeatedly attempted to delete a variable until it stopped. The - signs at the beginning of each row indicate removing a predictor. The output of the corresponding steps of the stepwise-selected model was returned with the following command and was shown in Table 6.10.

```
>model.s$anova
```

Table 6.10: The output of *model.s$anova*

Steps		Variables	Df	Deviance	Resid. Df	Resid. Dev	AIC
1			NA	NA	697	875.53	983.53
2	-	roadtypeAll_Trim3_nox	1	0.0014	698	875.53	981.53
3	-	no2_Trim3	1	0.0166	699	875.55	979.55
4	-	roadtype3_Trim1_nox	1	0.0178	700	875.57	977.57
5	-	fcc3_distance	1	0.0376	701	875.60	975.60
6	-	no2_Preg	1	0.0479	702	875.65	973.65
7	-	roadtypeAll_1stYr_nox	1	0.0506	703	875.70	971.70
8	-	no2_Trim1	1	0.0725	704	875.78	969.78
9	-	roadtype4_Trim3_nox	1	0.1086	705	875.88	967.88
10	-	pm10_Trim3	1	0.1307	706	876.01	966.01
11	-	roadtype3_Trim2_nox	1	0.1751	707	876.19	964.19
12	-	pm10_1stYr	1	0.1922	708	876.38	962.38
13	-	roadtype4_Trim1_nox	1	0.2324	709	876.61	960.61
14	-	roadtypeAll_2ndYr_nox	1	0.3246	710	876.94	958.94
.		.	.	.	.	.	.
.		.	.	.	.	.	.
.		.	.	.	.	.	.
21	-	roadtype4_Trim2_nox	1	0.1682	717	879.55	947.55
22	-	fcc4_distance	1	0.6116	718	880.16	946.16
23	-	no2_1stYr	1	0.6188	719	880.78	944.78
24	-	fcc1_distance	1	0.6882	720	881.46	943.46
25	-	no2_Trim2	1	0.7458	721	882.21	942.21
26	-	o3_Trim2	1	0.9486	722	883.16	941.16
27	-	pm10_Trim2	1	1.5242	723	884.68	940.68
28	-	o3_Trim3	1	1.0688	724	885.75	939.75
29	-	pm25_Trim2	1	1.3871	725	887.14	939.14
30	-	pm25_Trim1	1	1.1209	726	888.26	938.26
31	-	o3_Preg	1	1.4717	727	889.73	937.73
32	-	roadtypeAll_Trim2_nox	1	1.2244	728	890.96	936.96
33	-	pm10_Trim1	1	1.7838	729	892.74	936.74
34	-	pm25_1stYr	1	1.4421	730	894.18	936.18

From Table 6.10, it was realized that the AIC value decreased in each iteration due to the removal of an independent variable from the model based on the minimum AIC criteria. Finally, when the pm10_Trim1 variable was dropped from the model, the min. AIC=936.7398 value was obtained. As a result, at the last step, step() function determined the optimal set of features, that is, the *model.s* was achieved with fewer parameters and the minimum AIC value shown in Table 6.11.

```
> summary(model.s)
```

Table 6.11: Output of *summary(model.s)*

	Call:					
Regression Equation	*glm(formula = Autistic ~ fcc2_distance + roadtype1_1stYr_nox + roadtype1_2ndYr_nox + roadtype1_Preg_nox + roadtype1_Trim1_nox + roadtype1_Trim2_nox + roadtype1_Trim3_nox + roadtype3_1stYr_nox + roadtype3_Preg_nox + roadtype3_Trim3_nox + roadtype4_1stYr_nox + roadtype4_2ndYr_nox + roadtypeAll_Preg_nox + roadtypeAll_Trim1_nox + no2_2ndYr + o3_1stYr + o3_2ndYr + o3_Trim1 + pm25_2ndYr + gender, family = binomial(), data = exposure)*					
	Deviance Residuals: *Min 1Q Median 3Q Max* *-2.3355 -1.1613 0.6668 0.9682 1.6829*					
Coefficients	**Variables**	**Est. Coef.**	**Std.Error**	**z value**	**Pr(>|z|)**	**Signif. Codes**
	(Intercept)	-0.9049	0.7350	-1.2310	0.2183	
	fcc2_distance	0.0000	0.0000	-2.2320	0.0256	*
	roadtype1_1stYr_nox	0.1738	0.0519	3.3500	0.0008	***
	roadtype1_2ndYr_nox	-0.1382	0.0487	-2.8400	0.0045	**
	roadtype1_Preg_nox	-0.1024	0.0353	-2.9050	0.0037	**
	roadtype1_Trim1_nox	-0.0664	0.0232	-2.8660	0.0042	**
	roadtype1_Trim2_nox	0.0373	0.0239	1.5600	0.1188	
	roadtype1_Trim3_nox	0.0568	0.0266	2.1310	0.0331	*
	roadtype3_1stYr_nox	0.1005	0.0392	2.5600	0.0105	*
	roadtype3_Preg_nox	-0.0739	0.0368	-2.0080	0.0446	*
	roadtype3_Trim3_nox	-0.0911	0.0340	-2.6770	0.0074	**
	roadtype4_1stYr_nox	0.2568	0.1343	1.9120	0.0559	.
	roadtype4_2ndYr_nox	-0.2034	0.1362	-1.4940	0.1352	
	roadtypeAll_Preg_nox	0.0234	0.0140	1.6740	0.0942	.
	roadtypeAll_Trim1_nox	0.0364	0.0143	2.5540	0.0106	*
	no2_2ndYr	-0.0668	0.0355	-1.8830	0.0597	.
	o3_1stYr	-0.0761	0.0295	-2.5750	0.0100	*
	o3_2ndYr	0.0671	0.0296	2.2660	0.0234	*
	o3_Trim1	0.0128	0.0081	1.5740	0.1156	
	pm25_2ndYr	0.1269	0.0398	3.1900	0.0014	**
	genderMALE	0.3791	0.2194	1.7280	0.0840	.
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 983.31 on 750 degrees of freedom Residual deviance: 894.18 on 730 degrees of freedom AIC: 936.18					

As reported by Table 6.11, after the implementation of the *step()* function, *model.s* had the lowest AIC value among all the possible models and finally, 20 independent variables remained in the *model.s*. Additionally, *roadtype1_Trim2_nox, o3_Trim1 roadtype4_2ndYr_nox,* variables were significant between 0.15 and 0.1 level and *roadtypeAll_Preg_nox, gender* variables were significant between 0.05 and 0.1 level. On the other hand, other variables were statistically significant at the 5% level. When *roadtype1_Trim2_nox, roadtype4_2ndYr_nox and o3_Trim1, roadtypeAll_Preg_nox, and gender* variables were dropped from *model.s* and refitted, there was no improvement in the AIC value of the model. So, the *model.s* was assumed to be the final model.

The AIC values of the *model.2, model.p and model.s* were extracted below by using *model$aic* command. Unfortunately, they had the same AIC value. If they had different AIC values, the model with minimum AIC value would be preferred.

Extracting of model AIC Values

```
> model.s$aic
[1] 936.1654
> model.p$aic
[1] 936.1654
> model.2$aic
[1] 936.1654
```

6.3 Converting Continuous Fields to Categorical

In this study, to achieve interquartile change in odds ratio, the continuous fields in the “Exposure” table were converted to categorical variables in a new table, “*Exposure.q”*, with the following codes.

Converting Continuous Fields to Categorical

```
#Converting continuous fields to categorical selected by stepwise wrt
#AIC in (model.s)
library(dplyr)
#Selection the fields from model.s
test=summary(model.s)$coefficients
d=dim(test)
d=d-1  #last is gender
row=rownames(test)
fcolumns=""
for(i in 2:d[1]){
  col=row[i]
  if (i==2)
    fcolumns=paste(fcolumns,col,sep="")
  else
    fcolumns=paste(fcolumns,col,sep=" + ")
}

#Creating Exposure.q table
expformula=paste("select(Exposure,", fcolumns, ", gender, Autistic)")
Exposure.q=eval(parse(text = expformula))
str(Exposure.q)

#Adding new variables to the Exposure.q dataset for 19 continuous
variables
#Filling quartile value Q1=1, Q2=2
columns=""
for(i in 1:19){
   col=colnames(Exposure.q)[i]
   columns=paste(columns,col,sep="+")
   q=quantile(eval(parse(text=paste("Exposure.q$", col, sep = ""))))
#Obtain Quartiles
   b=c(q)
   b[1]=-Inf
   colq=paste(col ,"_Q",sep="")
   Exposure.q[,colq]=cut(eval(parse(text=paste("Exposure.q$", col, sep =
""))),breaks=b,labels=c(1,2,3,4))
}
str(Exposure.q)

#Getting the quartile variables for the formula in regression
fcolumns=""
for(i in 1:19){
  col=colnames(Exposure.q)[i]
  colq=paste(col ,"_Q",sep="")
  if (i==1)
    fcolumns=paste(fcolumns,colq,sep="")
  else
    fcolumns=paste(fcolumns,colq,sep=" + ")
}
#Forming model.q
glmformula=paste("glm(Autistic  ~ ", fcolumns, "+
gender,data=Exposure.q,family=binomial)")
model.q=eval(parse(text = glmformula))
summary(model.q)
```

In this code, 19 new categorical variables ending with "_Q" label for each existing continuous variables were added to the "*Exposure.q"* table. The new categorical variables show the quartiles of variables coded as 1,2,3, or 4 to indicate which quarter the record belongs to. For example, if the quartile is first quartile (*Q1*) the value is 1 and so on. Then, the new model, *model.q,* with categorized variables was fitted by using "*Exposure.q"* data with the following command.

Fitting of model.q	*>model.q =glm(formula = Autistic ~ fcc2_distance_Q + roadtype1_1stYr_nox_Q + roadtype1_2ndYr_nox_Q + roadtype1_Preg_nox_Q + roadtype1_Trim1_nox_Q + roadtype1_Trim2_nox_Q + no2_2ndYr_Q roadtype1_Trim3_nox_Q + roadtype3_1stYr_nox_Q + o3_1stYr_Q + roadtype3_Preg_nox_Q + roadtype3_Trim3_nox_Q + roadtype4_1stYr_nox_Q + roadtype4_2ndYr_nox_Q + roadtypeAll_Preg_nox_Q + o3_2ndYr_Q roadtypeAll_Trim1_nox_Q gender + o3_Trim1_Q + pm25_2ndYr_Q, family = binomial(), data = Exposure.step)*

Next, the details of the fitted model, *model.q,* were obtained with the *summary(model.q)* command, and the results were shown in Table 6.12.

```
> summary(model.q)
```

Table 6.12: Summary results of the fitting *model.q*

	Call:					
Regression Equation	*glm(formula = Autistic ~ fcc2_distance_Q + roadtype1_1stYr_nox_Q + roadtype1_2ndYr_nox_Q + roadtype1_Preg_nox_Q + roadtype1_Trim1_nox_Q + roadtype1_Trim2_nox_Q + roadtype1_Trim3_nox_Q + roadtype3_1stYr_nox_Q + roadtype3_Preg_nox_Q + roadtype3_Trim3_nox_Q + roadtype4_1stYr_nox_Q + roadtype4_2ndYr_nox_Q + roadtypeAll_Preg_nox_Q + roadtypeAll_Trim1_nox_Q + no2_2ndYr_Q + o3_1stYr_Q + gender + o3_2ndYr_Q + o3_Trim1_Q + pm25_2ndYr_Q, family = binomial(), data = Exposure.step)*					
	Deviance Residuals: *Min 1Q Median 3Q Max* *-2.6480 -1.1622 0.6483 0.9211 1.7962*					
Coefficients	**Variables**	**Est. Coef.**	**Std.Error**	**z value**	**Pr(>|z|)**	**Signif. Codes**
	(Intercept)	-0.5342	0.4166	-1.2820	0.1998	
	fcc2_distance_Q2	0.0099	0.2478	0.0400	0.9681	
	fcc2_distance_Q3	-0.1624	0.2506	-0.6480	0.5169	
	fcc2_distance_Q4	-0.2876	0.2495	-1.1530	0.2491	
	roadtype1_1stYr_nox_Q2	1.5495	0.8991	1.7230	0.0848	.
	roadtype1_1stYr_nox_Q3	1.5500	0.9984	1.5530	0.1205	
	roadtype1_1stYr_nox_Q4	2.6985	1.0221	2.6400	0.0083	**
	roadtype1_2ndYr_nox_Q2	-1.7478	0.8588	-2.0350	0.0418	*
	roadtype1_2ndYr_nox_Q3	-1.7931	0.9783	-1.8330	0.0668	.
	roadtype1_2ndYr_nox_Q4	-2.0521	0.9825	-2.0890	0.0367	*
	roadtype1_Preg_nox_Q2	0.7515	0.5914	1.2710	0.2038	

	roadtype1_Preg_nox_Q3	0.0338	0.7046	0.0480	0.9617	
	roadtype1_Preg_nox_Q4	-0.5536	0.7696	-0.7190	0.4719	
	roadtype1_Trim1_nox_Q2	-0.7491	0.4261	-1.7580	0.0788	.
	roadtype1_Trim1_nox_Q3	-0.9009	0.5222	-1.7250	0.0845	.
	roadtype1_Trim1_nox_Q4	-1.3193	0.5390	-2.4480	0.0144	*
	roadtype1_Trim2_nox_Q2	-0.3378	0.4800	-0.7040	0.4816	
	roadtype1_Trim2_nox_Q3	0.1490	0.6088	0.2450	0.8067	
	roadtype1_Trim2_nox_Q4	0.4337	0.6370	0.6810	0.4960	
	roadtype1_Trim3_nox_Q2	-0.1071	0.3709	-0.2890	0.7728	
	roadtype1_Trim3_nox_Q3	0.2345	0.4752	0.4930	0.6217	
	roadtype1_Trim3_nox_Q4	0.0454	0.5064	0.0900	0.9286	
	roadtype3_1stYr_nox_Q2	0.1104	0.3557	0.3100	0.7563	
	roadtype3_1stYr_nox_Q3	0.3418	0.4115	0.8310	0.4062	
	roadtype3_1stYr_nox_Q4	0.5841	0.4512	1.2940	0.1955	
	roadtype3_Preg_nox_Q2	-0.0432	0.3898	-0.1110	0.9117	
	roadtype3_Preg_nox_Q3	-0.0553	0.4804	-0.1150	0.9083	
	roadtype3_Preg_nox_Q4	-0.3220	0.5353	-0.6020	0.5475	
	roadtype3_Trim3_nox_Q2	-0.2980	0.3360	-0.8870	0.3752	
	roadtype3_Trim3_nox_Q3	-0.6795	0.4228	-1.6070	0.1080	
	roadtype3_Trim3_nox_Q4	-1.1524	0.4909	-2.3480	0.0189	*
	roadtype4_1stYr_nox_Q2	0.4627	0.3745	1.2350	0.2167	
	roadtype4_1stYr_nox_Q3	0.5205	0.4693	1.1090	0.2673	
	roadtype4_1stYr_nox_Q4	0.4141	0.5635	0.7350	0.4624	
	roadtype4_2ndYr_nox_Q2	-0.2786	0.3780	-0.7370	0.4611	
	roadtype4_2ndYr_nox_Q3	-0.3796	0.4654	-0.8160	0.4148	
	roadtype4_2ndYr_nox_Q4	0.0253	0.5661	0.0450	0.9643	
	roadtypeAll_Preg_nox_Q2	-0.1422	0.3637	-0.3910	0.6959	
	roadtypeAll_Preg_nox_Q3	0.2345	0.4941	0.4750	0.6351	
	roadtypeAll_Preg_nox_Q4	-0.0869	0.5546	-0.1570	0.8754	
	roadtypeAll_Trim1_nox_Q2	0.8721	0.3375	2.5840	0.0098	**
	roadtypeAll_Trim1_nox_Q3	1.3770	0.4536	3.0360	0.0024	**
	roadtypeAll_Trim1_nox_Q4	1.6898	0.5192	3.2540	0.0011	**
	no2_2ndYr_Q2	-0.2456	0.2425	-1.0130	0.3112	
	no2_2ndYr_Q3	0.0068	0.2835	0.0240	0.9808	
	no2_2ndYr_Q4	-0.4059	0.2822	-1.4380	0.1504	
	o3_1stYr_Q2	-0.1441	0.3102	-0.4650	0.6422	
	o3_1stYr_Q3	-0.1888	0.3642	-0.5180	0.6042	
	o3_1stYr_Q4	-0.3202	0.4140	-0.7740	0.4392	
	genderMALE	0.3971	0.2243	1.7700	0.0767	.
	o3_2ndYr_Q2	-0.0447	0.3133	-0.1430	0.8866	
	o3_2ndYr_Q3	-0.0863	0.3632	-0.2380	0.8122	
	o3_2ndYr_Q4	-0.0245	0.4045	-0.0610	0.9516	
	o3_Trim1_Q2	0.6613	0.2437	2.7140	0.0066	**
	o3_Trim1_Q3	0.4195	0.2582	1.6250	0.1042	
	o3_Trim1_Q4	0.7337	0.2826	2.5970	0.0094	**
	pm25_2ndYr_Q2	0.7831	0.2341	3.3450	0.0008	***
	pm25_2ndYr_Q3	0.8722	0.2804	3.1100	0.0019	**
	pm25_2ndYr_Q4	0.9418	0.2994	3.1460	0.0017	**
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 983.31 on 750 degrees of freedom Residual deviance: 888.76 on 692 degrees of freedom AIC: 1006.8					

For Table 6.12, the dummy variables representing the reference levels of categorical variables were omitted and the difference of each quartile level to the reference level was shown with the estimated coefficients. The z and p-values here denoted whether the difference of each level of the categorical variable to this reference level differed from zero or not. For example, *roadtype1_1stYr_nox_Q* categorical variable had four levels coded as 1, 2, 3, or 4 to show quarters (*Q1, Q2, Q3,or Q4*) of the r*oadtype1_1stYr_nox*. Then, three dummy variables (*roadtype1_1stYr_nox_Q2, roadtype1_1stYr_nox_Q3 and roadtype1_1stYr_nox_Q4)* were created and *roadtype1_1stYr_nox_Q1* was used as a reference level. The base level, *roadtype1_1stYr_nox_Q1*, was omitted and each coefficient of *roadtype1_1stYr_nox_Q* denoted the difference between the coefficient of *roadtype1_1stYr_nox_Q1* and the corresponding level. So the difference between the coefficient of *roadtype1_1stYr_nox_Q1* and *roadtype1_1stYr_nox_Q2* was 1.55. Another interpretation, in this case, could be: being in the second quartile(*Q2*) versus being first quartile(*Q1*) for *roadtype1_1stYr_nox_Q*, changes the log odds of being Autistic by 1.55 which was calculated by using Eq. 4.38 as

$$log\big(odds_{\mathrm{Q2}}\big) - log\big(odds_{\mathrm{Q1}}\big) = \log(\frac{odds_{\mathrm{Q2}}}{odds_{\mathrm{Q1}}}) = 1.55 \qquad (6.7)$$

Moreover, some of the levels of categorical variables were nonsignificant. Unless all non-reference levels were significant, we didn’t have to remove categorical values from the model. Some dummy variables were dropped by using stepwise elimination with help of *step()* function and a new model, *model.q.s,* with categorized variables was obtained from *model.q* for “*Exposure.q”* data shown below.

```
> model.q.s = step(model.q,direction="backward",test="Chisq",
trace=0)#trace=0 suppresses step by step output
```

Then, the details of the *model.q.s* were shown in Table 6.13 by using the *summary(model.q.s)* command.

```
> summary(model.q.s)
```

Table 6.13: Summary results of fitting *model.q.s*

	Call:					
Regression Equation	*glm(formula = Autistic ~ roadtype1_1stYr_nox_Q + roadtype1_2ndYr_nox_Q + roadtype1_Trim1_nox_Q + oadtype3_Trim3_nox_Q + roadtypeAll_Trim1_nox_Q + gender + o3_Trim1_Q + pm25_2ndYr_Q, family = binomial(), data = Exposure.q)*					
	Deviance Residuals: *Min 1Q Median 3Q Max* *-2.5260 -1.2046 0.7194 0.9325 1.5897*					
Coefficients	**Variables**	**Est. Coef.**	**Std.Error**	**z value**	**Pr(>|z|)**	**Signif. Codes**
	(Intercept)	-0.7526	0.3251	-2.3150	0.0206	*
	roadtype1_1stYr_nox_Q2	1.9377	0.8317	2.3300	0.0198	*
	roadtype1_1stYr_nox_Q3	1.7944	0.9358	1.9180	0.0552	.
	roadtype1_1stYr_nox_Q4	2.7387	0.9399	2.9140	0.0036	**
	roadtype1_2ndYr_nox_Q2	-1.8440	0.8163	-2.2590	0.0239	*
	roadtype1_2ndYr_nox_Q3	-1.7440	0.9303	-1.8750	0.0608	.
	roadtype1_2ndYr_nox_Q4	-2.0195	0.9227	-2.1890	0.0286	*
	roadtype1_Trim1_nox_Q2	-0.6186	0.3454	-1.7910	0.0733	.
	roadtype1_Trim1_nox_Q3	-0.8238	0.4206	-1.9590	0.0501	.
	roadtype1_Trim1_nox_Q4	-1.3764	0.4350	-3.1640	0.0016	**
	roadtype3_Trim3_nox_Q2	-0.1379	0.2316	-0.5950	0.5515	
	roadtype3_Trim3_nox_Q3	-0.3746	0.2630	-1.4240	0.1543	
	roadtype3_Trim3_nox_Q4	-0.6932	0.2766	-2.5060	0.0122	*
	roadtypeAll_Trim1_nox_Q2	0.8201	0.2875	2.8520	0.0043	**
	roadtypeAll_Trim1_nox_Q3	1.3413	0.3680	3.6450	0.0003	***
	roadtypeAll_Trim1_nox_Q4	1.6103	0.4007	4.0190	0.0001	***
	genderMALE	0.3460	0.2155	1.6060	0.1083	
	o3_Trim1_Q2	0.7168	0.2320	3.0900	0.0020	**
	o3_Trim1_Q3	0.3856	0.2348	1.6420	0.1005	
	o3_Trim1_Q4	0.5482	0.2425	2.2610	0.0238	*
	pm25_2ndYr_Q2	0.6592	0.2130	3.0940	0.0020	**
	pm25_2ndYr_Q3	0.8371	0.2358	3.5500	0.0004	***
	pm25_2ndYr_Q4	0.7328	0.2347	3.1230	0.0018	**
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family taken to be 1) Null deviance: 983.31 on 750 degrees of freedom Residual deviance: 913.32 on 728 degrees of freedom AIC: 959.32					

According to the results of the fitting *model.q.s* in Table 6.13, it had 22 dummy variables. They were significant at the level of 0.15 except for r*oadtype3_Trim3_nox_Q2.* Since the other Q3 and Q4 quartiles of *roadtype3_Trim3_nox_Q* were significant, there was no need to drop it. Then, *model.q* and *model.q.s* comparison was performed by the LRT to check the overall effect of the dropped variables with the following code as

Likelihood Ratio Test

```
#Likelihood ratio test
> lrtest(model.q, model.q.s)

  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  59 -444.38
2  23 -456.66 -36 24.564     0.9256
```

For the above LRT result:

Since $\chi^2 = 24.56 > \chi^2{}_{(1-0.05,2)} = 5.99$, or $P(\geq \chi^2) = 0.93$, which was insignificant at the α = 0.05 level, the hypothesis H_0 was accepted according to Eq. 4.45a. Therefore, we concluded that the *model.q.s* was better than *model.q*. As we see from the following results, all the variables in *model.q.s* were significant at the 15% level.

6.4 Calculation Odds Ratios

6.4.1 Calculation of the Odds ratios for *model.s* step in Table 6.11

First of all, as a continuation of Chapter 6.1.4, since to explain the odds ratios are easier than log-odds as a measure of association, the odds ratios for *model.s* in Table 6.11 were computed by using Eq. 4.40, that is, by taking the exponent of the estimated coefficients. Then, the 95% confidence intervals for odds ratios were also obtained from Eq. 4.48. The OR values and their CI for *model.s* were shown in Table 6.14. The odds ratios and CI were obtained with the help of *exp()* and *coef()* functions in R :

```
> exp(cbind(OR=coef(model.s),confint(model.s)))
```

Table 6.14: The OR values and their CI for *model.s*

Variables	exp(β)	CI		ASD Risk
	OR	2.50%	97.50%	
fcc2_distance	1.0000	1.0000	1.0000	(-,+)
roadtype1_1stYr_nox	**1.1898**	1.0807	1.3275	(+)
roadtype1_2ndYr_nox	0.8709	0.7856	0.9539	(-)
roadtype1_Preg_nox	0.9026	0.8409	0.9662	(-)
roadtype1_Trim1_nox	0.9358	0.8924	0.9778	(-)
roadtype1_Trim2_nox	**1.0380**	0.9917	1.0896	(+)
roadtype1_Trim3_nox	**1.0584**	1.0052	1.1161	(+)
roadtype3_1stYr_nox	**1.1057**	1.0257	1.1966	(+)
roadtype3_Preg_nox	0.9288	0.8632	0.9982	(-)
roadtype3_Trim3_nox	0.9127	0.8531	0.9752	(-)
roadtype4_1stYr_nox	**1.2934**	0.9935	1.6873	(+)
roadtype4_2ndYr_nox	0.8156	0.6232	1.0665	(-)
roadtypeAll_Preg_nox	**1.0237**	0.9970	1.0535	(+)
roadtypeAll_Trim1_nox	**1.0372**	1.0095	1.0677	(+)
no2_2ndYr	0.9353	0.8722	1.0026	(-)
o3_1stYr	0.9267	0.8738	0.9814	(-)
o3_2ndYr	**1.0693**	1.0096	1.1344	(+)
o3_Trim1	**1.0129**	0.9970	1.0293	(+)
pm25_2ndYr	**1.1353**	1.0511	1.2287	(+)
genderMALE	**1.4615**	0.9488	2.2463	(+)

From Table 6.14, it was concluded that the atributes showing an association with autim risk were *roadtype1_1stYr_nox*, *roadtype1_Trim2_nox*, *roadtype1_Trim3_nox*, *o3_2ndYr*, *roadtype3_1stYr_nox*, *roadtype4_1stYr_nox*, *roadtypeAll_Preg_nox*, *o3_Trim1*, *roadtypeAll_Trim1_nox*, *pm25_2ndYr*, *gender*. The strength of associations between them was determined as weak (OR=1.1-1.5).

6.4.2 Calculation the OR for Per IQR

Recall from Eq. 4.40 and Eq. 4.42, e^{β_i} represents the change in the odds for per unit change in a single variable X_i holding other variables fixed where β_i is the regression coefficient of the *i*th independent variable. On the other hand, if the variable X_i increase from *m* to *m+IQR* then the OR, changes from e^{β_i} to $(e^{\beta_i})^{IQR}$. The odds ratios (*ORs*) for per IQR change were calculated by using Eq. 4.42 with the following code and results were shown in Table 6.15.

Calculation of the OR for per IQR change

```
#Calculation of the OR for per IQR change.
sink("ModelStep-OR-Out.txt")
test=summary(model.s)$coefficients
d=dim(test)
row=rownames(test)
cat("Coefficients","\t", "Estimate","\t", "IQR","\t","OR", "\n")
for(i in 2:d[1]){
  col=row[i]
  q=IQR(eval(parse(text=paste("Exposure.step$", col, sep = "")))) #Obtain
interQuartiles
  q=round(q,2)
  if (col=="genderMALE") q=1
  OR=round(exp(test[[i,1]]*q),8)
  b=round(test[[i,1]],8)
  #cat(col,"\t", test[[i,1]], "\n")
  cat(col,"\t", b,"\t", q,"\t",OR, "\n")
}
sink()
```

Table 6.15: Estimated ORs for per IQR change

Variables	**Estimated Coefficients β_i**	**IQR**	**ORs per IQR $(e^{\beta_i})^{IQR}$**	**OR%**
fcc2_distance	-0.000020	15400.50	0.736687	-26.33
roadtype1_1stYr_nox	0.173812	7.12	**3.447138**	**244.71**
roadtype1_2ndYr_nox	-0.138285	7.07	0.376186	-62.38
roadtype1_Preg_nox	-0.102445	8.11	0.435687	-56.43
roadtype1_Trim1_nox	-0.066388	9.01	0.549826	-45.02
roadtype1_Trim2_nox	0.037325	8.35	**1.365698**	**36.57**
roadtype1_Trim3_nox	0.056748	7.84	**1.560337**	**56.03**
roadtype3_1stYr_nox	0.100522	6.78	**1.976922**	**97.69**
roadtype3_Preg_nox	-0.073852	7.64	0.568799	-43.12
roadtype3_Trim3_nox	-0.091304	6.99	0.528233	-47.18
roadtype4_1stYr_nox	0.257288	2.54	**1.922278**	**92.23**
roadtype4_2ndYr_nox	-0.203832	2.43	0.609381	-39.06
roadtypeAll_Preg_nox	0.023410	17.64	**1.511273**	**51.13**
roadtypeAll_Trim1_nox	0.036479	18.41	**1.957336**	**95.73**
no2_2ndYr	-0.066897	3.57	0.787555	-21.24
o3_1stYr	-0.076073	8.00	0.544122	-45.59
o3_2ndYr	0.067000	8.00	**1.709161**	**70.92**
o3_Trim1	0.012858	20.00	**1.293260**	**29.33**
pm25_2ndYr	0.126861	3.19	**1.498834**	**49.88**
genderMALE	0.379432	1.00	**1.461454**	**46.15**

According to the results of Table 6.15, the attribute associations with autism were shown in Table 6.16 and the change in *OR* as percentage groups per IQR was also shown in Table 6.17.

Table 6.16: Strength of associations with Autism

Associations with Autism		
Weak	**Moderate**	**Strong**
roadtype1_Trim2_nox	roadtype1_Trim3_nox	roadtype1_1stYr_nox
o3_Trim1	roadtype3_1stYr_nox	
pm25_2ndYr	roadtype4_1stYr_nox	
genderMALE	roadtypeAll_Preg_nox	
	roadtypeAll_Trim1_nox	
	o3_2ndYr	

Table 6.17: The OR% groups for a per IQR change

OR% Groups			
25-50%	**51-75**	**76-100**	**>100**
o3_Trim1	roadtypeAll_Preg_nox	roadtype3_1stYr_nox	roadtype1_1stYr_nox
pm25_2ndYr	roadtype1_Trim3_nox	roadtype4_1stYr_nox	
genderMALE	o3_2ndYr	roadtypeAll_Trim1_nox	

6.4.3 Calculation the Odds Ratios (*ORs*) and CI's for *model.q.s*

Finally, the odds ratios and their CI's for *model.q.s* were computed for *model.q.s* and shown in Table 6.18 by using the following R command.

```
> exp(cbind(OR= coef(model.q.s),confint(model.q.s)))
```

Table 6.18: The OR values and their CI for *model.q.s*

Variables	exp(β)	CI	
	OR	**2.50%**	**97.50%**
(Intercept)	0.471141	0.248334	0.889837
roadtype1_1stYr_nox_Q2	**6.942932**	1.597018	48.739778
roadtype1_1stYr_nox_Q3	**6.016133**	1.106084	49.158934
roadtype1_1stYr_nox_Q4	**15.467140**	2.889108	128.432779
roadtype1_2ndYr_nox_Q2	0.158185	0.023032	0.665033
roadtype1_2ndYr_nox_Q3	0.174812	0.021572	0.938410
roadtype1_2ndYr_nox_Q4	0.132718	0.016409	0.683895
roadtype1_Trim1_nox_Q2	0.538692	0.270902	1.053089
roadtype1_Trim1_nox_Q3	0.438766	0.190893	0.995615
roadtype1_Trim1_nox_Q4	0.252491	0.106259	0.586482
roadtype3_Trim3_nox_Q2	0.871183	0.552134	1.370214
roadtype3_Trim3_nox_Q3	0.687567	0.409904	1.150836
roadtype3_Trim3_nox_Q4	0.499987	0.289430	0.856900
roadtypeAll_Trim1_nox_Q2	**2.270620**	1.301075	4.026636
roadtypeAll_Trim1_nox_Q3	**3.824173**	1.877816	7.963581
roadtypeAll_Trim1_nox_Q4	**5.004539**	2.308097	11.126732
genderMALE	**1.413435**	0.924135	2.154207
o3_Trim1_Q2	**2.047824**	1.304258	3.241591
o3_Trim1_Q3	**1.470500**	0.929812	2.336642
o3_Trim1_Q4	**1.730129**	1.078086	2.791871
pm25_2ndYr_Q2	**1.933275**	1.276056	2.943944
pm25_2ndYr_Q3	**2.309603**	1.460452	3.684652
pm25_2ndYr_Q4	**2.080850**	1.318790	3.312791

The OR values for *model.q.s* in Table 6.18 might be explained as odds being in the *j*th quartile (*Qj*) versus odds being reference quartile (*Q1*) which was set to unity for any categorical variable, was calculated by using Eq. 4.40 as

$$\mathrm{OR} = \frac{odds_{\mathrm{Qj}}}{odds_{\mathrm{Q1}}} = \mathrm{e}^{\beta_{\mathrm{j}}} \qquad (6.8)$$

For example, for the OR values in Table 6.16, odds being in the *2nd* quartile(*Q2*) versus odds being first(referent) quartile (*Q1*) was 6.94 and odds being in the *4*th quartile (*Q4*) versus odds being first quartile (Q1) was 15.47 for roadtype1_1stYr_nox_Q. Put another way, taking more NO_2 as in the *4*th quartile, odds of autism risk increased 15.47 times relative to the first quartile.

$$OR = \frac{Odds\ of\ Autism\ when\ taking\ NO2\ in\ Quartile{=}4}{Odds\ of\ Autism\ when\ taking\ NO2\ in\ Quartile{=}1} = \frac{15.47}{1} = 15.47 \qquad (6.9)$$

As we see from Table 6.18, the risk of autism increased in the second, third, and fourth quartiles due to the first (referent) quartile for *roadtype1_1stYr_nox, roadtypeAll_Trim1_nox, o3_ Trim1, pm25_2ndYr.*

6.4 Model Evaluation

6.4.1 Pearson Chi-squared GOF Test

Pearson Chi-squared GOF test for *model.2* was evaluated with the following R code:

Pearson Chi-squared GOF Test

```
#Pearson Chi-squared GOF Test
peares= residuals(model.2, type='pearson')
pearsum=sum(peares*peares)
df <-model.2$df.residual
pval <- pchisq(pearsum, df, lower.tail =F)
cat("Pearson Chi-squared GOF Test\n","Chi2","\t", "df","\t", "p-
value","\n",round(pearsum,2),"\t", df,"\t", round(pval,2))

Pearson Chi-squared GOF Test
Chi2      df     p-value
737.93  730   0.41
```

The Pearson Chi-squared GOF test indicated that we could accept the null hypothesis, H_0, and the model was well-fitted with a p-value=0.41 > 0.05 where H_0: the model M_0 fits vs. H_A: the model M_0 does not fit.

6.4.2 Deviance GOF Test

Deviance and Pearson GOF Statistics are quite similar to each other.
R code for creating deviance GOF test for *model.2* was evaluated with the following R code:

Deviance GOF Test

```
#Deviance GOF Test
devres<- residuals(model.2, type='deviance')
devsum=sum(devres*devres)
df=model.2$df.residual
pval <- pchisq(devsum, df, lower.tail =F)
cat("Deviance GOF Test\n","Chi2","\t", "df","\t", "p-value","\n",round(devsum,2),"\t", df,"\t",
round(pval,2))
Deviance GOF Test
Chi2       df    p-value
894.17   730   2.8e-05
```

The Deviance GOF test indicated that we could reject the null hypothesis, H_0, and the model was not well-fitted with a p-value=2.8 10-5< 0.05. where H_0: the model fits vs. H_1: the model does not fit. Since deviance was too large and or p-value was too small, the model didn't fit all the features in the data.

The R summary function, *summary(model.2)*, provided both null and residual deviance statistics with their degrees of freedom as:

summary (model.2)

```
     Null deviance: 983.31  on 750  degrees of freedom
Residual deviance: 894.17  on 730  degrees of freedom
AIC: 936.17
```

When we looked at *Residual deviance* in *summary(model.2)*, we noted that the *Residual deviance* results were identical (894.17) with Deviance GOF Test.

6.4.3 Hosmer-Lemeshow GOF Test

Hosmer-Lemeshow test was performed by using *logitgof(obs=, exp, g = 10)* function shown below for *model.2* where the arguments *obs* is a vector of observed values, *exp* is expected values fitted by the model and *g* number of groups.

Hosmer- Lemesshow GOF Test

```
#Hosmer- Lemesshow GOF Test
library("generalhoslem")
logitgof(Exposure$Autistic, fitted(model.2),g=10, ord = FALSE)
Hosmer and Lemeshow test (binary model)
data:  Exposure$Autistic, fitted(model.2)
X-squared = 11.826, df = 8, p-value = 0.1591
```

Hosmer- Lemeshow GOF Test indicated that we could accept the null hypothesis, H_0, and the model was well-fitted with a p-value=*0.1591* > 0.05. where H_0: observed-predicted=0, the model fits vs. H_1: observed-predicted≠0, the model does not fit.

7. SUMMARY OF RESULTS

In my research, the relation between four air pollutants (NO_2, O_3, PM_{10}, $PM_{2.5}$) and ASD for 6 time periods (all pregnancy, first trimester, second trimester, third trimester, first year and second year) were examined by using logistic regression models and variable selection methods.

Firstly, a reduced model, *model.2*, containing fewer variables was achieved from the full model, *model.full*, containing all variables by using purposeful variable selection. Then, *model.p* was obtained from *model.full* by the stepwise selection method according to *p-values*. Next, *model.s* was formed by using an automatic stepwise selection according to the AIC criterion. All the models (*model.2*, *model.p*, *model.s*) had the same AIC value and fewer predictors, approximately 20 predictors, than *model.full*. The variables in these models had *p-values* < 0.15 and were statistically significant at the 15% level.

With respect to the variable selection results, the fitting of reduced models, *model.2*, *model.p*, and *model.s* resulted in the same variable as shown in Table 7.1. For the Wald test results, 15 variables were found to be important and significant at the level of 0.05.

Table 7.1: The significant or important variables w.r.t. Wald test

Variables	Variable Explanation	Est. Coef.	Pr(>|z|)
fcc2_distance	Distance to nearest state highways	0.0000	0.0256
roadtype1_1stYr_nox	NO2 from Interstate highways during 1st year	0.1738	0.0008
roadtype1_2ndYr_nox	NO2 from Interstate highways during 2nd year	-0.1382	0.0045
roadtype1_Preg_nox	NO2 from Interstate highways during pregnancy	-0.1024	0.0037
roadtype1_Trim1_nox	NO2 from Interstate highways during 1st trimester	-0.0664	0.0042
roadtype1_Trim3_nox	NO2 from Interstate highways during 3rd trimester	0.0568	0.0331
roadtype3_1stYr_nox	NO2 from country highways during 1st year	0.1005	0.0105
roadtype3_Preg_nox	NO2 from country highways during pregnancy	-0.0739	0.0446
roadtype3_Trim3_nox	NO2 from country highways during 3rd trimester	-0.0911	0.0074
roadtype4_1stYr_nox	NO2 from city street during 1st year	0.2568	0.0559
roadtypeAll_Trim1_nox	NO2 from all roads during 1st trimester	0.0364	0.0106
no2_2ndYr	NO2 during 2nd year	-0.0668	0.0597
o3_1stYr	O3 during 1st year	-0.0761	0.0100
o3_2ndYr	O3 during 2nd year	0.0671	0.0234
pm25_2ndYr	Pm25 during 2nd year	0.1269	0.0014

Secondly, the odds ratios for *model.s* were calculated, and results shown in Table 7.2 The variables with odds ratios > 1 were associated with ASD risk. The ORs for NO$_2$ from interstate highways during the first year, NO$_2$ from interstate highways during the second trimester, NO$_2$ from interstate highways during the third trimester, NO$_2$ from country highways during the first year, NO$_2$ from interstate highways during the third trimester, NO$_2$ from country highways during the first year, NO$_2$ from city street during the first year, NO$_2$ from all roads during pregnancy, O$_3$ during the second year, O$_3$ during the trimester, Pm$_{2.5}$ during the second year were found to be 1.1898, 1.0380, 1.0584, 1.1057, 1.2934, 1.0237, 1.0372, 1.0693, 1.0129, 1.1353, 1.4615 respectively.

Table 7.2: The variables associated with ASD risk

Variables	Variable Explanation	exp(β)	ASD Risk
		OR	
fcc2_distance	Distance to nearest state highways	1.0000	(-,+)
roadtype1_1stYr_nox	NO_2 from Interstate highways during 1st year	1.1898	(+)
roadtype1_Trim2_nox	NO_2 from Interstate highways during 2nd trimester	1.0380	(+)
roadtype1_Trim3_nox	NO_2 from Interstate highways during 3rd trimester	1.0584	(+)
roadtype3_1stYr_nox	NO_2 from country highways during 1st year	1.1057	(+)
roadtype4_1stYr_nox	NO_2 from city street during 1st year	1.2934	(+)
roadtypeAll_Preg_nox	NO_2 from all roads during pregnancy	1.0237	(+)
roadtypeAll_Trim1_nox	NO_2 from all roads during 1st trimester	1.0372	(+)
o3_2ndYr	O_3 during 2nd year	1.0693	(+)
o3_Trim1	O_3 during trimester	1.0129	(+)
pm25_2ndYr	Pm25 during 2nd year	1.1353	(+)
genderMALE	Gender	1.4615	(+)

Then, per IQR change, the OR were calculated for *model.s,* and results were shown in Table 7.3. The results demonstrated that NO_2 from interstate highways during the first year was strongly associated with ASD, NO_2 (from interstate highways during the third trimester; from country highways during the first year; from city street during the first year; from all roads during the first trimester), and O_3 during the second year were moderately associated with ASD, and weakly associated with exposure to NO_2 from interstate highways during the second trimester, O_3 during the first trimester and $PM_{2.5}$ during the second year. They also showed drastic increases in OR% per IQR. Especially NO_2 (from interstate highways during the first year; from country highways during the first year; from city street during the first year; from all roads during the first trimester) increased more than 90%, NO_2 (from interstate highways during the third trimester; from all roads during pregnancy), and O_3 during the second year increased more than 50%, and the others increased less than 50%.

Table 7.3 The *ORs* for each change in *IQR* for *model.s*

Variables	Exposure Explanation	IQR	ORs per IQR Change	OR% per IQR	St. of Ass. with ASD
roadtype1_1stYr_nox	NO_2 from Interstate highways during 1st year	7.1	3.45	244.71	S
roadtype1_Trim2_nox	NO_2 from Interstate highways during 2nd trimester	8.4	1.37	36.57	W
roadtype1_Trim3_nox	NO_2 from Interstate highways during 3rd trimester	7.8	1.56	56.03	M
roadtype3_1stYr_nox	NO_2 from country highways during 1st year	6.8	1.98	97.69	M
roadtype4_1stYr_nox	NO_2 from city street during 1st year	2.5	1.92	92.23	M
roadtypeAll_Preg_nox	NO_2 from all roads during pregnancy	17.6	1.51	51.13	M
roadtypeAll_Trim1_nox	NO_2 from all roads during 1st trimester	18.4	1.96	95.73	M
o3_2ndYr	O_3 during 2nd year	8.0	1.71	70.92	M
o3_Trim1	O_3 during 1st trimester	20.0	1.29	29.33	W
pm25_2ndYr	$Pm_{2.5}$ during 2nd year	3.2	1.50	49.88	W
genderMALE	Gender	1.0	1.46	46.15	W
Strength of Association with ASD: W=Weak (OR=1 - 1.5), M=Moderate (OR=1.51-2.5) and S=Strong (OR>2.5)					

Additionally, to achieve interquartile change in odds ratio, the continuous fields in my data were converted into categorical variables with four levels coded as 1, 2, 3, or 4 to show quarters (*Q1, Q2, Q3, or Q4*). Then, by using stepwise elimination, a new reduced model, *model.q.s,* was obtained from the *model.q c*ontaining categorized variables. Next, the *OR* were calculated for that model and the OR values > 1 were shown in Table 7.3. The results showed that the relative increase in the odds of ASD, from Q1 to Q2 was 6.94, and 15.47 from Q1 to Q4, when exposed to NO_2 in Q1 from interstate highways during 1st year, going from Q1 to Q2, is 2.27, from Q1to Q4 was 5.00 when exposed to NO_2 in Q1 from all roads during the first trimester and so on. Finally, Pearson Chi-squared, Deviance, and Hosmer-Lemeshow GOF Tests were used. to evaluate the model and to assess the GOF, According to Pearson Chi-squared and Hosmer-Lemeshow GOF Tests, the *model.2* was well-fitted.

Table 7.4: The OR values > 1, the relative increase in the odds of ASD from one level of quartile to another, for *model.q.s*

Variables	Explanation	exp(β)
		OR
roadtype1_1stYr_nox_Q2	Taken NO2 in Q2 from Interstate highways during 1st year	6.94
roadtype1_1stYr_nox_Q3	Taken NO2 in Q3 from Interstate highways during 1st year	6.02
roadtype1_1stYr_nox_Q4	Taken NO2 in Q4 from Interstate highways during 1st year	15.47
roadtypeAll_Trim1_nox_Q2	Taken NO2 in Q2 NO2 from all roads during 1st trimester	2.27
roadtypeAll_Trim1_nox_Q3	Taken NO2 in Q3 NO2 from all roads during 1st trimester	3.82
roadtypeAll_Trim1_nox_Q4	Taken NO2 in Q4 NO2 from all roads during 1st trimester	5.00
genderMALE	genderMALE	1.41
o3_Trim1_Q2	Taken NO2 in Q2 O3 during 1st trimester	2.05
o3_Trim1_Q3	Taken NO2 in Q3 O3 during 1st trimester	1.47
o3_Trim1_Q4	Taken NO2 in Q4 O3 during 1st trimester	1.73
pm25_2ndYr_Q2	Taken NO2 in Q2 Pm25 during 2nd year	1.93
pm25_2ndYr_Q3	Taken NO2 in Q3 Pm25 during 2nd year	2.31
pm25_2ndYr_Q4	Taken NO2 in Q4 Pm25 during 2nd year	2.08

8. CONCLUSIONS

Firstly, the ORs for one unit increase in NO_2 exposure were found to be 1.19 from interstate highways during the first year, 1.04 from interstate highways during the second trimester, 1.06 from interstate highways during the third trimester, 1.11 from country highways during the first year, 1.29 from city street during the first year, 1.02 from all roads during pregnancy and 1.04 from all roads during the first trimester. Then, the ORs for Ozone (O_3) exposure during the second year and during trimester were 1.07 and 1.01 respectively. Finally, the ORs for Ozone (O_3) exposure during the second year was (OR=1.14).

Secondly, when we examined adjusted odds ratios (AOR) for ASD per IQR increase in NO_2, O_3 and $PM_{2.5}$ exposures from different roads during different periods according to Table 7.3, we detected elevated AORs (OR=3.44 per 7.1 ppb [IQR] increase in NO_2 from interstate highways during first year; OR=1.98 per 6.8 ppb [IQR] increase in NO_2 from country highways during first year; OR=1.96 per 18.4 ppb [IQR] increase in NO_2 from all roads during first trimester; OR=1.92 per 2.5 ppb [IQR] increase in NO_2 from city street during first year; OR=1.71 per 8 ppb [IQR] increase in O_3 during the second year; OR=1.56 per 7.8 ppb [IQR] increase in NO_2 from interstate highways during the third trimester; OR=1.51 per 17.6 ppb [IQR] increase in NO_2 from all roads during pregnancy; OR=1.46 per 3.2 ppb [IQR] increase in $PM_{2.5}$ during the second year; OR=1.36 per 8.4 ppb [IQR] increase in $PM_{2.5}$ from interstate highways during second trimester; OR=1.29 per 20 ppb [IQR] increase in O_3 during first trimester).

Finally, when the highest level of NO_2, O_3 and $PM_{2.5}$ exposures were compared to lowest level due to Table 7.4; Subjects exposed to NO_2 from interstate highways during the first year, exposed to NO_2 in Q4 from all roads during the first trimester, exposed to O_3 in Q4 during the first trimester, and exposed to PM_{25} in Q4 during the second year, were associated with ASD risk (OR=15.47), (OR=5) (OR=1.73) and (OR=2.08) respectively compared to lowest quartile (Q1).

REFERENCES

[1] Volk, H., Hertz-Picciotto, I., Delwiche, L., Lurmann, F. and McConnell, R. (2011). Residential Proximity to Freeways and Autism in the CHARGE Study. Environmental Health Perspectives, 119(6), pp.873-877.

[2] Volk, H., Lurmann, F., Penfold, B., Hertz-Picciotto, I. and McConnell, R. (2013). Traffic-Related Air Pollution, Particulate Matter, and Autism. JAMA Psychiatry, 70(1), p.71.

[3] Gong, T., Almqvist, C., Bölte, S., Lichtenstein, P., Anckarsäter, H., Lind, T., Lundholm, C. and Pershagen, G. (2014). Exposure to Air Pollution From Traffic and Neurodevelopmental Disorders in Swedish Twins. *Twin Research and Human Genetics*, 17(6), pp.553-562.

[4] Oudin, A., Frondelius, K., Haglund, N., Källén, K., Forsberg, B., Gustafsson, P. and Malmqvist, E. (2019). Prenatal exposure to air pollution as a potential risk factor for autism and ADHD. Environment International, 133, p.105149.

[5] Centers for Disease Control and Prevention. 2020. *Research | Autism Spectrum Disorder (ASD) | CDC*. [online] Available at: <https://www.cdc.gov/ncbddd/autism/research.html#ref> [Accessed 24 July 2020].

[6] Ninds.nih.gov. 2020. *Autism Spectrum Disorder Fact Sheet | National Institute Of Neurological Disorders And Stroke*. [online] Available at: <https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Autism-Spectrum-Disorder-Fact-Sheet#3082_5> [Accessed 24 July 2020].

[7] Al-Hamdan, A., Preetha, P., Albashaireh, R., Al-Hamdan, M. and Crosson, W. (2018). Investigating the effects of environmental factors on autism spectrum disorder in the USA using remotely sensed data. *Environmental Science and Pollution Research*, 25(8), pp.7924-7936.

[8] Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C. M., & Reichenberg, A. (2014). The familial risk of autism. JAMA, 311(17), 1770–1777.

[9] Yang, C., Zhao, W., Deng, K., Zhou, V., Zhou, X. and Hou, Y. (2017). The association between air pollutants and autism spectrum disorders. *Environmental Science and Pollution Research*, 24(19), pp.15949-15958.

[10] Black, D., 2015. DSM-5® Guidebook: The Essential Companion to the Diagnostic and Statistical Manual of Mental Disorders. 5th edn. Edited by

Donald W. Black and Jon E. Grant (567 pp., ISBN 9781585624652). American Psychiatric Association Publishing.

[11] Faculty.washington.edu. 2020. *Neuroscience For Kids - Brain Development*. [online] Available at: <https://faculty.washington.edu/chudler/dev.html> [Accessed 25 July 2020].

[12] Centers for Disease Control and Prevention. 2020. *Early Brain Development And Health | CDC*. [online] Available at: <https://www.cdc.gov/ncbddd/childdevelopment/early-brain-development.html> [Accessed 25 July 2020].

[13] Larose, D. and Larose, C., 2014. *Discovering Knowledge In Data*- An Introduction to Data Mining

[14] https://www.apa.org. 2020. *Autism And Autism Spectrum Disorders*. [online] Available at: <https://www.apa.org/topics/autism/> [Accessed 26 July 2020].

[15] Exkorn, K., Foreword by Volkmar, Fred R., M.D., 2005. *The Autism Sourcebook.* PerfectBound Publisher.

[16] Baio, J., Wiggins, L., Christensen, D., Maenner, M., Daniels, J., Warren, Z., Kurzius- Spencer, M., Zahorodny, W., Robinson, C., Rosenberg, White, T., Durkin, M., Imm, P., Nikolaou, L., Yeargin-Allsopp, M., Lee, L., Harrington, R., Lopez, M., Fitzgerald, R., Hewitt, A., Pettygrove, S., Constantino, J., Vehorn, A., Shenouda, J., Hall-Lande, J., Van, K., Naarden, Braun and Dowling, N. (2018). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR. Surveillance Summaries,* 67(6), pp.1-23.

[17] Baio J, Wiggins L, Christensen DL, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. MMWR Surveill Summ 2018;67(No. SS-6):1– 23.

[18] Nimh.nih.gov. 2020. *NIMH » Autism Spectrum Disorder (ASD)*. [online] Available at: <https://www.nimh.nih.gov/health/statistics/autism-spectrum-disorder-asd.shtml> [Accessed 25 July 2020].

[19] Centers for Disease Control and Prevention. 2020. *Data And Statistics On Autism Spectrum Disorder | CDC*. [online] Available at: <https://www.cdc.gov/ncbddd/autism/data.html> [Accessed 25 July 2020].

[20] Sphweb.bumc.bu.edu. 2020. *Autism*. [online] Available at: <http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH/Autism/Autism_print.html> [Accessed 25 July 2020].

[21] Autism Reading Room. 2020. *Risk Factors - Autism Reading Room*. [online] Available at: <http://readingroom.mindspec.org/?page_id=892> [Accessed 25 July 2020].

[22] Ninds.nih.gov. 2020. *Autism Spectrum Disorder Fact Sheet | National Institute Of Neurological Disorders And Stroke*. [online] Available at: <https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Autism-Spectrum-Disorder-Fact-Sheet#3082_5> [Accessed 24 July 2020].

[23] IACAPAP. 2020. *JM Rey's IACAPAP E-Textbook Of Child And Adolescent Mental Health - IACAPAP*. [online] Available at: <https://iacapap.org/iacapap-textbook-of-child-and-adolescent-mental-health/> [Accessed 24 July 2020].

[24] Centers for Disease Control and Prevention. 2020. *Data And Statistics On Autism Spectrum Disorder | CDC*. [online] Available at: <https://www.cdc.gov/ncbddd/autism/data.html#references> [Accessed 25 July 2020].

[25] Landrigan PJ (2010) What causes autism? Exploring the environmental contribution. Curr Opin Pediatr 22:219-225. PMID: 20087185.

[26] Glock, M., Glock, M., Glock, M., Glock, M. and Cacchiotti, N., 2020. *Autism Genetics & The Enviornment | Autism Research Institute*. [online] Autism Research Institute. Available at: <https://www.autism.org/genetics-the-environment-and-autism> [Accessed 25 July 2020].

[27] Lyall K, Schmidt RJ, Hertz-Picciotto I. Maternal lifestyle and environmental risk factors for autism spectrum disorders. Int J Epidemiol. 2014;43(2):443–464.

[28] Carautismroadmap.org. 2020. *What Causes Autism? | Center For Autism Research*. [online] Available at: <https://www.carautismroadmap.org/what-causes-autism/> [Accessed 25 July 2020].

[29] Anna Oudin, Kasper Frondelius, Nils Haglund, Karin Källén, Bertil Forsberg, Peik Gustafsson, Ebba Malmqvist, Prenatal exposure to air pollution as a potential risk factor for autism and ADHD, Environment International,Volume 133, Part A, 2019,105149,ISSN 0160-4120.

[30] Reference, G., 2020. *ASD*. [online] Genetics Home Reference. Available at: <https://ghr.nlm.nih.gov/condition/autism-spectrum-disorder#genes> [Accessed 25 July 2020].

[31] National Institute of Environmental Health Sciences. 2020. *Autism*. [online] Available at: <https://www.niehs.nih.gov/health/topics/conditions/autism/index.cfm#footnote7 > [Accessed 25 July 2020].

[32] Tox Town. 2020. *Air Pollution: Your Environment, Your Health | National Library Of Medicine*. [online] Available at: <https://toxtown.nlm.nih.gov/sources- of-exposure/air-pollution> [Accessed 25 July 2020].

[33] US EPA. 2020. *Air Topics | US EPA*. [online] Available at: <https://www.epa.gov/environmental-topics/air-topics> [Accessed 25 July 2020].

[34] Raz, R., Levine, H., Pinto, O., Broday, D., Yuval and Weisskopf, M. (2017). Traffic-Related Air Pollution and Autism Spectrum Disorder: A Population-Based Nested Case-Control Study in Israel. American Journal of Epidemiology, 187(4), pp.717-725.

[35] Www3.epa.gov. 2020. *Pollutants And Sources | Technology Transfer Network Air Toxics Web Site | US EPA*. [online] Available at: <https://www3.epa.gov/ttn/atw/pollsour.html> [Accessed 25 July 2020].

[36] Tox Town. 2020. *Air Pollution: Your Environment, Your Health | National Library Of Medicine*. [online] Available at: <https://toxtown.nlm.nih.gov/sources- of-exposure/air-pollution> [Accessed 25 July 2020].

[37] Becerra, T., Wilhelm, M., Olsen, J., Cockburn, M. and Ritz, B. (2013). Ambient Air Pollution and Autism in Los Angeles County, California. *Environmental Health Perspectives*, 121(3), pp.380-386.

[38] 4cleanair.org. 2020. *Air Pollutants | Public Site*. [online] Available at: <http://www.4cleanair.org/topics/details/air-pollutants> [Accessed 25 July 2020].

[39] Raz, R., Roberts, A., Lyall, K., Hart, J., Just, A., Laden, F. and Weisskopf, M. (2015). Autism Spectrum Disorder and Particulate Matter Air Pollution before, during, and after Pregnancy: A Nested Case–Control Analysis within the Nurses' Health Study II Cohort. *Environmental Health Perspectives*, 123(3), pp.264-270.

[40] Ntp.niehs.nih.gov. 2020. *Traffic-Related Air Pollution And Hypertensive Disorders Of Pregnancy*. [online] Available at: <https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/completed/pollution/index.html> [Accessed 25 July 2020].

[41] https://www.epa.gov/transportation-air-pollution-and-climate-change/smog-soot-and-local-air-pollution

[42] Tox Town. 2020. *Nitrogen Oxides: Your Environment, Your Health | National Library Of Medicine*. [online] Available at: <https://toxtown.nlm.nih.gov/chemicals-and-contaminants/nitrogen-oxides> [Accessed 25 July 2020].

[43] US EPA. 2020. *Basic Information About NO2 | US EPA*. [online] Available at: <https://www.epa.gov/no2-pollution/basic-information-about-no2#What%20is%20NO2> [Accessed 25 July 2020].

[44] Ritz, B., Liew, Z., Yan, Q., Cuia, X., Virk, J., Ketzel, M. and Raaschou-Nielsen, O. (2018). Air pollution and autism in Denmark. Environmental Epidemiology, 2(4), p.e028.

[45] Gong, T., Dalman, C., Wicks, S., Dal, H., Magnusson, C., Lundholm, C., Almqvist, C. and Pershagen, G. (2017). Perinatal Exposure to Traffic-Related Air Pollution and Autism Spectrum Disorders.

[46] Pagalan, L., Bickford, C., Weikum, W., Lanphear, B., Brauer, M., Lanphear, N., Hanley, G., Oberlander, T. and Winters, M. (2019). Association of Prenatal Exposure to Air Pollution With Autism Spectrum Disorder. *JAMA Pediatrics*, 173(1), p.86.

[47] Tox Town. 2020. *Particulate Matter: Your Environment, Your Health | National Library Of Medicine*. [online] Available at: <https://toxtown.nlm.nih.gov/chemicals-and-contaminants/particulate-matter> [Accessed 25 July 2020].

[48] US EPA. 2020. *Particulate Matter (PM) Basics | US EPA*. [online] Available at: <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM> [Accessed 25 July 2020].

[49] Talbott, E., Arena, V., Rager, J., Clougherty, J., Michanowicz, D., Sharma, R. and Stacy, S. (2015). Fine particulate matter and the risk of autism spectrum disorder. *Environmental Research*, 140, pp.414-420.

[50] Chen, G., Jin, Z., Li, S., Jin, X., Tong, S., Liu, S., Yang, Y., Huang, H. and Guo, Y. (2018). Early life exposure to particulate matter air pollution (PM1, PM2.5

and PM10) and autism in Shanghai, China: A case-control study. *Environment International*, 121, pp.1121-1127.

[51] Geng, R., Fang, S. and Li, G. (2019). The association between particulate matter 2.5 exposure and children with autism spectrum disorder. *International Journal of Developmental Neuroscience*, 75, pp.59-63.

[52] Jo, H., Eckel, S., Wang, X., Chen, J., Cockburn, M., Martinez, M., Chow, T., Molshatzki, N., Lurmann, F., Funk, W., Xiang, A. and McConnell, R. (2019). Sex-specific associations of autism spectrum disorder with residential air pollution exposure in a large Southern California pregnancy cohort. *Environmental Pollution*, 254, p.113010.

[53] Guxens, M., Ghassabian, A., Gong, T., Garcia-Esteban, R., Porta, D., Giorgis-Allemand, L., Almqvist, C., Aranbarri, A., Beelen, R., Badaloni, C., Cesaroni, G., de Nazelle, A., Estarlich, M., Forastiere, F., Forns, J., Gehring, U., Ibarluzea, J., Jaddoe, V., Korek, M., Lichtenstein, P., Nieuwenhuijsen, M., Rebagliato, M., Slama, R., Tiemeier, H., Verhulst, F., Volk, H., Pershagen, G., Brunekreef, B. and Sunyer, J. (2016). Air Pollution Exposure during Pregnancy and Childhood Autistic Traits in Four European Population-Based Cohort Studies: The ESCAPE Project. Environmental Health Perspectives, 124(1), pp.133-140.

[54] Pagalan, L., Bickford, C., Weikum, W., Lanphear, B., Brauer, M., Lanphear, N., Hanley, G., Oberlander, T. and Winters, M. (2019). Association of Prenatal Exposure to Air Pollution With Autism Spectrum Disorder. *JAMA Pediatrics*, 173(1), p.86.

[55] Kalkbrenner, A., Windham, G., Serre, M., Akita, Y., Wang, X., Hoffman, K., Thayer, B. and Daniels, J. (2015). Particulate Matter Exposure, Prenatal and Postnatal Windows of Susceptibility, and Autism Spectrum Disorders. Epidemiology, 26(1), pp.30-42.

[56] Kim, D., Volk, H., Girirajan, S., Pendergrass, S., Hall, M., Verma, S., Schmidt, R., Hansen, R., Ghosh, D., Ludena-Rodriguez, Y., Kim, K., Ritchie, M., Hertz-Picciotto, I. and Selleck, S. (2017). The joint effect of air pollution exposure and copy number variation on risk for autism. Autism Research, 10(9), pp.1470-1480.

[57] Yousefian, F., Mahvi, A., Yunesian, M., Hassanvand, M., Kashani, H. and Amini, H. (2018). Long-term exposure to ambient air pollution and autism spectrum disorder in children: A case-control study in Tehran, Iran. *Science of The Total Environment*, 643, pp.1216-1222.

[58] Tox Town. 2020. *Ozone: Your Environment, Your Health | National Library Of Medicine*. [online] Available at: <https://toxtown.nlm.nih.gov/chemicals-and- contaminants/ozone> [Accessed 25 July 2020].

[59] US EPA. 2020. *Ground-Level Ozone Basics | US EPA*. [online] Available at: <https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#wwh> [Accessed 25 July 2020].

[60] 4cleanair.org. 2020. *Ozone / Smog | Public Site*. [online] Available at: <http://www.4cleanair.org/topics/story/ozone-smog> [Accessed 25 July 2020].

[61] Jung, C., Lin, Y. and Hwang, B. (2013). Air Pollution and Newly Diagnostic Autism Spectrum Disorders: A Population-Based Cohort Study in Taiwan. PLoS ONE, 8(9), p.e75510.

[62] Kerin, T., Volk, H., Li, W., Lurmann, F., Eckel, S., McConnell, R. and Hertz-Picciotto, I. (2017). Association Between Air Pollution Exposure, Cognitive and Adaptive Function, and ASD Severity Among Children with Autism Spectrum Disorder. Journal of Autism and Developmental Disorders, 48(1), pp.137-150.

[63] Kaufman, J., Wright, J., Rice, G., Connolly, N., Bowers, K. and Anixt, J. (2019). Ambient ozone and fine particulate matter exposures and autism spectrum disorder in metropolitan Cincinnati, Ohio. *Environmental Research*, 171, pp.218-227.

[64] Ndar.nih.gov. 2020. *NIMH Data Archive (NDA)*. [online] Available at: <https://ndar.nih.gov/index.html> [Accessed 25 July 2020].

[65] Nda.nih.gov. 2020. *NIMH Data Archive - Data Dictionary*. [online] Available at: <https://nda.nih.gov/data_dictionary.html?source=NDA&submission=ALL> [Accessed 25 July 2020].

[66] Han, J., Kamber, M. and Pei, J., 2012. *Data Mining*. Waltham, Mass.: Morgan Kaufmann Publishers.

[67] Guide to Intelligent Data Analysis.How to Intelligently Make Sense of Real Data. Michael R. Berthold, Christian Borgelt, Frank Höppner, Frank Klawonn (2010). Springer-Verlag London Limited

[68] Data Mining Practical Machine Learning Tools and Techniques (3rd Ed)-Ian H.

[69] Tan, P., Steinbach, M., Karpatne, A. and Kumar, V., 2014. *Introduction To Data Mining*.

[70] Exploration, A., 2020. *A Complete Tutorial Which Teaches Data Exploration In Detail*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/> [Accessed 25 July 2020].

[71] Tukey, JW. Exploratory data analysis. Addison-Wesely, 1977

[72] Itl.nist.gov. 2020. *Box Plot*. [online] Available at: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda337.htm> [Accessed 25 July 2020].

[73] Stapel, E., 2020. *Interquartile Ranges (Iqrs) & Outliers | Purplemath*. [online] Purplemath. Available at: <http://www.purplemath.com/modules/boxwhisk3.htm> [Accessed 25 July 2020].

[74] Harrell, F., 2015. Regression Modeling Strategies. 2nd ed. Switzerland: SPRINGER.

[75] David W. Hosmer (2013). *Applied Logistic Regression*, Third Edition- John Wiley & Sons, Inc.

[76] Kutner M. H., Nachtsheim C. J., Neter J., and Li W., 2005., Applied Linear Statistical Model, Firth Edition. McGraw-Hill

[77] Kleinbaum, D. G., & Klein, M. (2010). Logistic regression(statistics for biology and health) (3rd ed.). Springer-Verlag New York Inc.

[78] Agresti, A., 2019. An Introduction To Categorical Data Analysis. 3rd ed. Edition. JohnWiley & Sons, Inc

[79] Rosenthal, J., 2012. Statistics And Data Interpretation For Social Work. New York, NY: Springer.

[80] Jacob Cohen, Patricia Cohen, Stephen G. West, Leona S. Aiken (2003), *Applied Multiple Regression*/*Correlation Analysis* for the *Behavioral Sciences, 3rd Edition.* Lawrence Erlbaum Associates, Inc.

[81] Forthofer, R., 2007. Biostatistics A Guide To Design, Analysis, And Discovery. 2nd ed. England: ACADEMIC PR-ELSEVIER Science (MO).

[82] Wiki.ecdc.europa.eu. 2020. *Model Building Strategies*. [online] Available at: <https://wiki.ecdc.europa.eu/fem/Pages/Model%20building%20strategies.aspx> [Accessed 25 July 2020].

[83] Bilder, C. and Loughin, T., 2015. *Analysis Of Categorical Data With R*.

[84] Simon J. Sheather (2009), Modern Approach to Regression with R, Springer Science + Business Media

[85] Bennette, C. and Vickers, A., 2012. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology, 12(1).

[86] Forthofer, R., 2007. *Biostatistics A Guide To Design, Analysis, And Discovery*. 2nd ed. England: ACADEMIC PR-ELSEVIER Science (MO).

[87] Bilder, C. and Loughin, T., 2015. *Analysis Of Categorical Data With R*. Boca Raton: CRC Press Taylor & Francis Group.

[88] Hilbe, J., 2009. *Logistic Regression Models*. Boca Raton: Chapman & Hall/CRC.

[89] Crawley, M., 2013. The R Book. 2nd ed. Chichester, England: Wiley.

[90] Hilbe, J., 2015. *Practical Guide To Logistic Regression*. Boca Raton: CRC Press.

[91] Verzani, J., 2011. Getting Started With Rstudio. Sebastopol, CA: O'Reilly.

[92] R-project.org. 2020. *R: The R Project For Statistical Computing*. [online] Available at: <https://www.r-project.org/> [Accessed 25 July 2020].

[93] Williams, G., 2011. Data Mining With Rattle And R. New York, NY: Springer Science+Business Media, LLC.

[94] Kabacoff, R., 2011. R In Action: Data Analysis And Graphics With R. Manning Publications.

[95] Newton, I. (2014). Minitab cookbook. Birmingham, UK: Packt Publ.

[96] Lee, C., Lee, J., Chang, J. and Tai, T. (2016). Essentials of Excel, Excel VBA, SAS and Minitab for Statistical and Financial Analyses. Cham: Springer International Publishing.

[97] Faraway, J., 2015. *Linear Models With R*. 2nd ed.

CURRICULUM VITAE

Personal Information

Name Surname	: Tamer Demir
Place and Date of Birth	: Istanbul, 20.06.1965

Education

Undergraduate	: Istanbul Technical University, Physics Engineering, 1992-1995
Graduate	: Middle East Technical University, Physics Department, 1983-1987
Secondary	: Kuleli Military High School, Istanbul, 1979-1983
Foreign Language Skills	: English-Good

Work Experience

2010 - 2018	: Data Bilişim Hizmetleri (DBH) Software Development & *Consulting*
2001 - 2010	: Teknoloji Holding - Teknoser A.Ş, IT Manager
1996 - 2001	: Zeytinoğlu Holding , Computer Programmer & IT Chief
1994 - 1995	: Özel Bahçeşehir College, Computer Teacher & IT Manager
1987 - 1994	: T.C Land Forces (K.K.K) / Balıkesir OBI Chief Information Officer

Contact:

Telephone	: +90 532 769 59 23
E-mail Address	: tamer.demir34@gmail.com