



**ADDIS ABABA UNIVERSITY  
GRADUATE STUDIES PROGRAMME  
DEPARTMENT OF STATISTICS**

**Ordinal Logistic Regression Analysis of  
Correlates of Crime Severity:**  
*The case of Tigray Region, Ethiopia*

By

Meron Desalegn

**A Thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of Science in Statistics.**

**JUNE, 2011**

**ADDIS ABABA**

**ADDIS ABABA UNIVERSITY  
GRADUATE STUDIES PROGRAMME  
DEPARTMENT OF STATISTICS**

# **Ordinal Logistic Regression Analysis of Correlates of Crime Severity:**

*The case of Tigray Region, Ethiopia*

**A Thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of Science in Statistics.**

**By Meron Desalegn**

**Advisor: Dr.Grima Taye**

**JUNE, 2011  
ADDIS ABABA**

**ADDIS ABABA UNIVERSITY  
GRADUATE STUDIES PROGRAMME  
DEPARTMENT OF STATISTICS**

By Meron Desalegn

**Approved by the Board of Examiners:**

\_\_\_\_\_  
Chairman

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Examiner

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Examiner

\_\_\_\_\_  
Signature

---

Tel : +251-11-239465

P.O.BOX 1176

## *Acknowledgments*

*First, and foremost, I would like to thank the almighty God for giving me the opportunity to pursue my graduate study at Department of Statistics, Addis Ababa University.*

*I owe the deepest gratitude to Dr.Girma Taye, my thesis advisor for his valuable and constructive comments and encouragements throughout my study.*

*My deepest thanks go to my parents and sister, who have supported me all the way to fulfill my dream, and I am highly indebted to my friend Gizachew Kefelew for his unreserved support and encouragement throughout my study.*

*Finally I am also grateful to all my friends at the department of statistics especially, Ato Seifu Neda, Samual Muluye and Ketema Kebebew, for all encouragement and support that they gave me while I was writing this thesis.*

# TABLE OF CONTENTS

ACRONYMS.....	v
ABSTRACT.....	vi
<b>CHAPTER1:INTRODUCTION.....</b>	<b>1</b>
1.1Background .....	1
1.2 Statement of the Problem.....	7
1.3 Objectives of the Study.....	8
1.4 Significance of the Study.....	8
1.5 Limitations of the Study.....	9
<b>CHAPTER 2: LITERATURE REVIEW.....</b>	<b>11</b>
2.1 Theoretical Review of Literature.....	11
2.2 Empirical Review of Literature.....	15
<b>CHAPTER 3 :_DATA AND METHODOLOGY.....</b>	<b>18</b>
3.1 Sample Size Determination.....	18
3.2 Study Area.....	19
3.3 Data Source.....	20
3.4 Model.....	22
3.5 Ordinal Logistic Regression .....	24
3.6 Proportional Odds Model.....	24
3.7 Partial Proportional Odds Model .....	27
3.8 Odds Ratio.....	29
3.9 Model selection.....	29
3.10 Goodness-of-Fit Measures .....	30
3.10. Model Adequacy.....	33

<b>CHAPTER 4: RESULT AND DISCUSSION.....</b>	<b>36</b>
4.1 Descriptive statistics.....	36
4.2 Logistic Regression Analysis .....	37
4.3 Interpretation of the results.....	45
4.4 Discussion of the result.....	49
<b>CHAPTER 5: CONCLUSION AND RECOMMENDATION.....</b>	<b>52</b>
5.1 Conclusions .....	52
5.2 Recommendations.....	53
<b>REFERENCES.....</b>	<b>54</b>
ANNEX1: SAS out Put .....	58
A.1.1 Logistic regression Full Model.....	58
A.1.2 Logistic Regression Rediuced Model .....	60
A.1.3 Interaction Model fit for Rediuced Model.....	63
ANNEX 2: Assessing of Empty Cell .....	64
ANNEX 3: Descriptive Statistics.....	66
ANNEX 4: Graph : Test of Paralel and Linear Assumption.....	69
ANNEX 5: About Type 3 Analysis of SAS.....	71
ANNEX 6: SAS Code.....	72

## ACRONYMS

AIC	Akaike Information Criteria
EDHS	Demographic Health Survey
FDREPCC	Federal Democratic Republic Ethiopia Population and Census Commission
FMOHE	Federal Ministry of Health
GED	General Educational Development
GLM	General Linear Model
ICMA	International City Managers Association
LL	Log Likelihood
Logit	Log of Odds
ML	Maximum Likelihood
OLS	Ordinary Least Square
OR	Odds Ratio
PO	Proportional Odds
PPOM	Partial Proportional Odds Model
PPOM-UR	Unrestricted Partial Proportional Odds Model
UCR	Uniform Crime Report

## ABSTRACT

The cumulative logit or the proportional odds regression model is one of the popular choices to study covariate effects on ordinal responses. This research is aimed at modeling a categorical response i.e., crime severity outcome in terms of some predictors, determines the goodness of fit as well as validity of the assumptions and selecting an appropriate and more parsimonious model there by proffer useful suggestions and recommendations. The proportional odds model was used as a tool to model the two major factors viz. socio-demographic (sex, age, education status, marital and employment status) and environmental (urban, rural) that affected the outcomes of crime. The fit, of the model was illustrated with 2,753 crime records obtained from regional police commission. This study provides some graphical and numerical methods for checking the adequacy of the proportional odds regression model. The methods focus on evaluating crime severity for specific covariate effects. The tested model showed good fit and performed differently depending on categorization of outcome, adequacy in relation to assumptions and goodness of fit. Findings of this study have shown that criminal's age, educational background, employment status, marital status, and area of crime committed are significantly affect the outcome of crime. Criminals who are young, illiterate, employed, single household member, and rural areas of crime committed increased the odds of being in either serious or medium crime categories, than those criminals who are aged, educated, unemployed, married, and urban area of crime committed. The finding of this study indicates that the rise of crime in the region is generally as the result of direct effect of poverty. Policies and plans have to be put in place to improve young age individual education and crime prevention agencies need to issue to implement crime prevention strategies.

**Keywords:** crime severity, ordinal logistic regression, proportional odds model.

# **CHAPTER ONE**

## **INTRODUCTION**

### **1.1 BACKGROUND**

Human beings need to live and work in a place where they are safe. They want to ensure that there is a concerned body that protects their lives as well as their properties from potential hazards. In fact, one of the main functions of any government is to ensure that law and order for the security of its citizens are put in place (Wilson, 1963). In other words, far from formulation and enactment of law for the prevention of crime, governments must establish agencies and organizations that enforce the same.

Accordingly, in many countries, there are organizations such as courts, prosecutions and police, which are responsible for the observance of law and order.

In Ethiopia, the police organization was established in 1942 under proclamation No.6 as autonomous institution with the responsibility of preventing and investigating crime incidents. In 1996, the police institution was put under the ministry of interior. Since its establishment, the police organization structure has been extended to the lower administration level namely “Woreda” and sometimes “kebele” (Mesfin, 1999).

Due to the new constitution adopted in 1994, the Ethiopian government has been exercising federal political system and hence both the structure and authority of the police has changed accordingly. Based on the proclamation No. 1 article 50, regional governments are duly responsible to establish all the necessary administrative levels in their respective region. In light with this political sphere, regional states established their own police institutions at a level of commission. As a result, the Tigray Police Commission has become responsible to maintain law and order in its respective region together with other concerned agencies.

Along with the prevention and investigation of crime, police makes use of previous crime reports and data as an input for the formulation of crime prevention policies and strategic plans (Wilson, 1993). It is obvious from the outset that to make use of data and records,

relevant data have to be kept and managed properly. For this reason, the Tigray police commission has been collecting criminal records since its establishment and has maintained numerous criminal records consisting of fingerprints, names, photographs and general descriptions of criminals.

Crime records, here, refers to the data gathered and used by the police other than the Administrative records of the police itself. According to the ICMA (1961), the classes of crime records can be divided into three broad categories. These are:

- *Cases or complaint records:* This includes information regarding complaints and reports received by the police from citizens and other agencies, and actions indicated by the police.
- *Arrest records:* This class contains all records about arrested offenders including their control and disposition. The scope of arrest records covers every step from the arrest of a person till s/he is released.
- *Personal Identification records:* This major division of police records consists of records dealing with personal identification of criminals.

This investigation adopted personal identification records, which include attributes of criminals such as name, age, sex, educational status, and occupation, ethnic background, address and so on.

### **Crime and Criminals**

A general statement often made is that there is no development where there is no peace. It is said that “African policy makers are keenly aware of the fact that substantial improvements in the economic and social situation of their population are contingent upon the maintenance of peace. Without peace, little or nothing can be achieved” (ECA, 2004:3). General reference is made, in the case of the African region, to armed conflicts, which are per-se violation of rule of law. It should be noted that armed conflicts are made accompanied of different types of crime and that crime is identified as one of the major challenges to development.

Thakur (2003) had defined crime as an act or omission of an act, which is punishable by law. However, an act that is considered as a crime in one place and time may not be true in another place or time. Consistent with their nature crimes are broadly categorized in two class: violent and property crime. Violent crimes are committed against people and include murder, rape, robbery, forcible and aggravated assault, the incidence of such crime is very high. Property crimes consist of burglary, larceny, motor vehicle theft, vandalism, check forgery, larceny-theft and arson.

According to Andargachew (1998), a criminal was defined as a person who has violated the legally forbidden acts. In fact, there are some factors that have to be taken in to account to convict whether a person should be considered as a criminal or not. Among these, an individual should be of competent age in light with the law of the land; and there must be a well-predefined punishment for the particular act committed. Sutherland and Cressey stated that an act would be considered as a crime if it is prohibited by the criminal law. Criminal law, on the other hand, refers to a body of specific rules regarding human conduct, which have been explicitly enacted by sovereign body.

Crime has increasingly become as complex as human nature. Technological advancement and tremendous progress in communication have facilitated criminals of every corner of the world to commit a crime using sophisticated equipment in one place and escape to another place (Thakur, 2003). Nowadays, the world is facing the proliferation of problems such as illicit drug trafficking, smuggling, hijacking, kidnapping, and terrorism.

Crime has adversely affected the society of both civilized as well as developing countries by declining the quality of life, endangering human right and fundamental freedom and posing a serious challenge to the community. Although the level and intensity of the problem might vary from nation to nation, no country has remained unaffected.

Regarding the seriousness of crimes, there is a remarkable agreement around the world. All major and minor types of crimes, burglary, robbery, assault and car theft, to mention just a few, are noticeable all over the world. Over a five year period, two out of three

inhabitants of big cities are victimized by crime at least once and the risk of being victimized are highest in Latin America and (sub Saharan) Africa. Globally, the chances to be victimized by serious crimes robbery, sexual crimes or assault are one in five<sup>1</sup>.

Empirical crime models have generally assumed criminals as a rational, and utility maximizing individuals so that decisions to commit a crime are based on cost benefit considerations. The costs and benefits are evaluated on individual utility functions which of course diverge across criminals. Many socioeconomic variables are also assumed to influence criminal behavior by affecting the costs and benefits of committing a crime.

As crime is one of the social evils, it has a direct impact on human wellbeing. For this reason, governments usually establish organizations such as prosecutorial, judicial, correctional and probation.

Thakur (2003) suggested that intent and opportunity are two major factors that lead to the occurrence of a crime. An individual cannot commit a crime unless and otherwise s/he gets an opportunity even though s/he has intended to commit a crime. Therefore, the best strategy for crime prevention is to provide a system that denies any opportunity for a criminal to commit a crime.

However, these days, law enforcement and investigating agencies have recognized the tremendous value in extracting hidden knowledge embedded in their data warehouses which could be valuable in the process of combating crimes (Megaputer Intelligence, 2002). The police departments want to reveal frequent crime patterns from historical reports to help them investigate new cases.

Using reported crimes to compile city and county crime rankings often leads to simplest and incomplete analysis that can create misleading perceptions which may adversely affect cities, counties, and their residents. Geographic and demographic factors to each

---

<sup>1</sup> [http://www.uncjin.org/special/global Report.html](http://www.uncjin.org/special/global%20Report.html)

jurisdiction must be considered and applied to make an accurate and complete assessment of crime in a given jurisdiction (Brown, 2003).

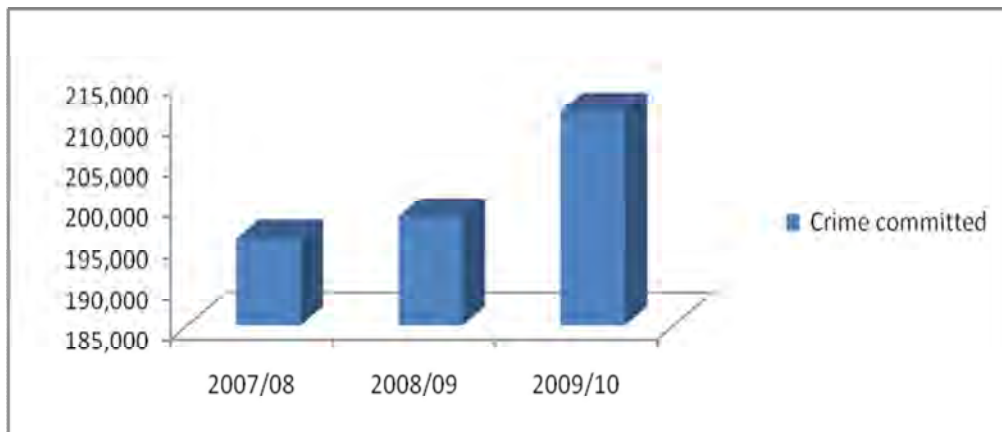
According to Megaputer Intelligence (2002), the analysis of crime patterns and trends is very important for police officers and analysts to learn from historical crime patterns and enhance crime resolution rate. It also helps to prevent future incidents by putting in place preventive mechanisms based on observed patterns. Another possible advantage is, it can reduce the training time for officers assigned to a new location and having no prior knowledge of site specific patterns to assist them in investigations. In light with the crime patterns extracted from previous records, police deploys scarce resources to the right place at the right time.

Two-third of victims of serious crimes who had reported their victimization to the police indicated an unmet need for help. Levels of demand for help among victims were highest in Central and Eastern Europe, Asia, Africa and Latin America. Reporting is particularly low in the countries of Latin America. World wide, less than half of the victims who reported their cases to the police were satisfied with the response. Levels of satisfaction with the police are lowest among victims in Latin America, Central and Eastern Europe and Africa. The most common crime reported was theft, followed by burglary. Violent crime (homicide, assaults, or robbery) had less share (around 10-15 percent) of all reported crime. The rates were higher for industrial countries than non-industrial countries. Arab States, generally, reported very low rates for nearly all types of crime. Cities around the world showed similar patterns for homicide and for robbery rates. However, high homicide rates were reported for several of the Latin American cities, New York and some northern European cities (<http://www.uncjin.org/special/globalReport.html>).

### **Crime Trend in Ethiopia**

In Ethiopia, crime statistics of the Federal Police Commission has revealed that the rate of crime is increasing steadily. Considering the three consecutive years in 2007/08 about 195,608 crimes were reported to the police excluding Somali region (federal police commission, 2010). Similarly, the report has publicized that in the year 2008/09 about

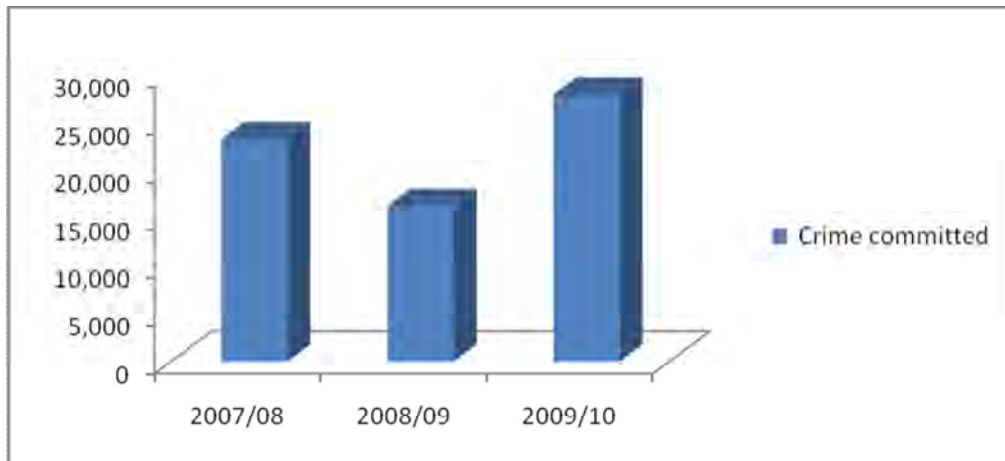
198,250 different crimes were reported excluding Harari region. The national crime statistics compiled by the Federal police commission in 2010/11 for the year 2009/10 has shown about 211,202 crimes were reported to police throughout the country with exception to SNNPR. A result indicated that number of crime reported in 2009/10 is increased by 6.5% and 7.4% as compared to the year 2007/08 and 2008/09, respectively. Figure 1 shows that the total number of crimes reported in the year 2009/10 has significant increment.



**Figure 1: The Ethiopian crime trend of three consecutive years.**

In the same way, about 23,232 crimes were reported throughout the Tigray region to the police in a given year of 2007/08, however 16,248 crimes were reported in year 2008/09 which indicates a significant decrement. The report also reveals that about 27,940 different crimes are committed in the year 2009/10.

Moreover, the national crime statistics report has indicated that the Tigray regional state accounts for 13% of the total crimes reported to the police in the entire country; and among individuals who commit crime the region took share of 12.8% in 2009/10. This indicates that the Tigray region is one of the crime prone regions in the country.



**Figure 2: Crime trend of Tigray region for three consecutive years.**

## 1.2 Statement of the Problem

Crime is a complex social phenomenon and its cost is increasing due to a number of societal changes and the like, and hence, law enforcement organizations like that of police need to learn the factors that contribute to the high crime trends (Wilson, 1963). To curb this social evil, there is always a need for prudent crime prevention strategies and policies. Understanding and processing criminal records is one method to learn about both crime and individuals who involve in misdeeds (Brown, 2003).

Tigray region has high population density, unemployment rate, poverty and other social problems, which may be the reason for occurrence of crime. This may be uncommon to the norms and culture, values and traditions of the population. Part of the solution should be giving attention to the issue of crime since it is becoming a transnational one and it can also be easily and rapidly spread or transferred across cultures.

Crimes are not committed randomly by individuals since the individuals take into account the cost benefit analysis of the same. There are complex circumstantial relationships between several characteristics (economical condition, stability of population, effective strength of law enforcement agencies, etc) and the incidence of crime. As such one cannot improve safety without successes fully reduced related crime rate and severity to

the causative variables. In addition to human life and bodily harm costs, the severity of the situation in economic terms is also very alarming.

The prevailing situation calls for intervention in view of minimizing the number and magnitude of crimes in different aspects. To this effect, there is a need for identifying major factors that cause the increment of crime rate. Accordingly, this study focuses on identifying factors that most explain the rate and occurrence of crime in the region.

### **1.5 Objective of the Study**

#### **General objective**

The main objective of this study is to demonstrate logistic analysis of crime data and identifying factors that influence the incidence of crime in Tigray region of Ethiopia.

#### **Specific objectives**

- To show the applications of ordinal logistic regression models.
- To explore the nature and incidence of crime at different levels.
- To assess the effect of multiple covariates in different levels of crime.

### **1.6 Significance of the Study**

This study attempted to assess crime patterns as related to criminal attributes and hence police officials can make use of these patterns in their day-to-day battle against crime.

Police officials in the crime prevention and investigation authority of the respective region can make use of the results of this study in order to make optimal deployment of resource in crime prevention. Moreover, the output of the study shall be used for designing appropriate training programs and crime prevention, investigation strategies, and this result helps as a basis for further study in this area.

The result of this study provides information to regional government and other concerned bodies to adjust their intervention programs, in setting additional policies, strategies, and further investigations.

The study will help both the government and non-governmental organizations who work on the prevention and control of crime to take continuous considerations of data collection on crime and factors associated with it, for further monitoring and evaluation assessment.

Model developed during this research will be used for determining and predicting occurrence of crime in a given space and time.

### **1.7 Limitations of the Study**

One of the major limitations of this study is related with the problem of the availability of complete crime data. Data available in police commission, court office, and justice office are recorded in different formats and registers; however, it is difficult to obtain full information on the crime variables especially from the data collection formats available in the early years. This may be due to the presence of less awareness on the proper data recording system. Nowadays, even though the data collection system is better, there are still problems since data on some of the crime variables are poorly documented with missing values. As a result, some covariates that have significant influence on the crime remain incomplete.

When a crime is committed there are individuals who are offenders of the crime, criminals and; victims of the crime. However, this study is limited to the study of criminal characteristics to extract crime trends. The case of victims is excluded from this study deliberately. The absence of sufficient literature with regard to the severity and occurrence of crime in our country is another limitation of the study.

# CHAPTER TWO

## LITERATURE REVIEW

### 2.1. Theoretical Literature Review

Criminology research has highlighted several important factors related to high crime rate including unemployment, economic backwardness, over population, illiteracy and inadequate equipment of the police force (Thakur, 2003). The form seriousness and size of the crime may rely on the form of a society and thus its nature changes with the growth and development of the social system. In every generation it has its own most critical, new and special problems of crime, although the crime problem is as old as man himself. In addition to this, the techniques employed to commit crime became modern in the sense that they make use of modern knowledge and technique. The rise in crime both at national and transitional is generally thought as the result of interplay between socio-economic changes. Consequently, there is a value system developed in those areas that tolerates delinquency and crime as normal parts of the residential life.

Wilson (1987) focused on the social and economic condition in the poor neighborhoods that are correlated with concentrated crime. His finding suggested that deteriorated social economic conditions and limited opportunities to get job create settings for all types of delinquency and crime as residents in the poor areas suffer from social and cultural isolation and inadequate social capital from mainstream culture. The interactions between criminal victimization and offending thus contribute to a vicious cycle of violence as residents with limited resources gradually make adaptations to deteriorated neighborhood conditions and become more tolerant of using violence as means of solution to daily problems (Sampson and Wilson, 1995; Andersons, 1999).

Other studies have shown that younger minority population of lower social economic status, are disproportionately involved in crime and are at highest risk of criminal offending and victimization. Sampson and Wilson (1995) highlighted a cycle of crime

that explains the high crime rate in poor inner city. The cycle begins when affluent residents together with industry and business move out of the inner city. This exodus leads to greater urban decay and other conditions of social disorganization including: a sharp increase in unemployed rate, people living in poverty, illegitimate births, single parent families, drug abuse and violence. With decline of economy in the neighborhoods, social institutions including churches, schools, stores, health care and recreational facilities, decline as well. This decaying process results in adequate informal social networks and creates opportunities for delinquency and crime (Wilson and Kelling, 1982). Other factors, such as residents' fear of crime, fear of encounter with the police and gang activities in the neighborhoods, make it more difficult to have an effective crime prevention and control.

In sum, previous researches of unemployment and crime suggested that social and economic factors create variations of crime concentration. The neighborhoods with concentrated poverty, high population mobility and inadequate informal social network will have a consistent concentration of crime over time.

Unemployment rate is one of the most widely referenced economic indicators. In discussions of potential impacts of the economy on crime rates, many scholars and policy makers use the unemployment rate as a proxy for economic strength. US congress report (1978) has shown interest in the relationship between the economies - unemployment, in particular and crime rates since the 1970s. The then recession of US, which has been accompanied by a rise in the unemployment rate, has once again focused attention on the relationship between unemployment and crime rates. A number of studies analyzing the relationship between unemployment and crime rates, in the nineteen-eighties, studies found that higher unemployment was associated with greater occurrence of crime; however the link was statistically looser than the link between measures of deterrence. "It failed to show a well-defined, quantifiable linkage" (Freeman, 1983).

There is a relatively large literature that link wages and unemployment rates to criminal behavior. Recent studies concluded that crime is increasing in local unemployment rates

and decreasing in wage rates. To the extent that education increases wage rates and reduces the likelihood of unemployment, it increases the opportunity costs of crime and will tend to reduce post-school criminal activity. Higher wages raise the opportunity costs of crime in two distinct ways. First, since crime may require time to commit, the time cannot be used for other productive purposes like work. Here, it is useful to think of all of the time involved in planning a crime, locating a target and, potentially, evading detection and arrest. Second, each crime committed entails an expected period of incarceration, which is more costly for individuals with better labor market opportunities and wages.

On one hand, property crimes like burglary, auto theft, and drug dealing can involve significant planning or time spent on the actual activity itself. On the other hand, violent crimes like assault would appear to require less time for planning and execution but are associated with higher expected probabilities of arrest, conviction, and incarceration as well as longer sentence lengths conditional on incarceration. Lochner (2004) calculated that for each assault, the perpetrator can expect to spend 63 days incarcerated; however, the expected incarceration period for a burglary is only 13 days. These time costs would appear to exceed the direct time costs associated with committing most crimes. Thus, changes in wages or Unemployment rates could have greater effects on violent crimes than on property crimes.

Education may also affect the rewards from crime. This is most likely to be true for white collar crimes like fraud, forgery, and embezzlement. Education may actually increase these types of crime if it increases the rewards from crime more than it increases legitimate wages. Lochner (2004) found some evidence that white collar crime rates are increasing in average education levels. To the extent that schools 'socialize' students to become better citizens and to treat others better, education may also reduce the psychic returns to crime causing individuals to forego lucrative criminal opportunities. Education may also teach individuals to be more patient. This will discourage crime, since forward-looking individuals place greater weight on any expected punishment associated with their criminal activities. Education may also affect preferences toward risk. To the extent that schooling makes individuals more risk averse, it will tend to discourage crime.

Empirically, there is a strong negative correlation between educational attainment and various measures of crime. Freeman (1996) pointed out that more than two-thirds of all incarcerated men in 1993 had not graduated from high school. The empirical literature on education and crime has focused almost exclusively on the effects of educational attainment on post-school criminal activity; however, a few studies have attempted to estimate the 'effects' of school enrollment on contemporaneous crime. Recently, few studies have attempted to estimate the effects of youth arrest and incarceration on educational outcomes. Studies which attempt to estimate the effects of law enforcement policies or criminal opportunities on educational decisions are virtually non-existent.

Developmental (age-based) theory suggests that different factors may have different effects at different ages when a person begins to commit crime whether a person continues to commit crime or stops. That the impact of age on criminal involvement is one of the strongest factors associated with crime has prompted the controversial claim that the age-crime relationship is universal, invariant, inexplicable with social science variables, and involves no interaction between age and any variable that explains or correlates with crime (Hirschi and Gottfredson, 1983).

The F.B.I.'s Uniform Crime Report (2009) data, particularly the Crime Index (homicide, robbery, rape, aggravated assault, burglary, larceny-theft, auto theft) document the robustness of the age effect on crime and also reveal a long-term trend toward younger age-crime distributions in more modern times. Crime has links to poverty, employment and housing conditions, as well as other factors such as the number of young people in the population and the availability of illegal drugs. City residents must deal with higher levels of crimes of violence and against their property, in addition to the fear that criminal activity instills. Today, the peak age (the age group with the highest age-specific arrest rate) is younger. Although crime tends to decline with age, substantial variation can be found in the parameters of the age-crime curve (such as peak age, median age, and rate of decline from peak age). "Flatter" age curves (i.e., those with an older peak age) and/or a

slower decline in offending rates among older age groups) are associated with several circumstances.

Tseloni (1999) compares single-level and multilevel negative binomial models of personal victimization counts. The two levels of analysis are used for the covariates which refer to person (level-1) include socio- demographic characteristics, such as sex, age, race, marital status, educational level employment status, and life style factor. The level-2 or household covariates include indicators of affluence, such as number of cars owned, annual family income, tenure and type of accommodation, namely number of adults and children in the household; and protection against crime. The negative binomial regression models of this study explicitly account for population unexplained heterogeneity which is estimated by respective coefficients of overdispersion. While both the NBM and the MNBM estimates the between individuals such heterogeneity the latter accounts for additional unexplained heterogeneity could be allocated over various sources of clustering, such as the segment, the census track or the state. The similarity of the estimated coefficients and their standard errors between the conventional and a multilevel specification indicates that the accounting for the clustering of sampling units does not influence the estimated fixed effects of personal crime covariates. However the two models offer a different story with regard to estimated unexplained heterogeneity between individuals.

Osgood (2000) introduced the use of regression models based on the Poisson distribution as a tool for resolving common problems in analyzing aggregate crime rates. When the population size of an aggregate unit is small relative to the offense rate, crime rates must be computed from a small number of offenses. Such data are ill-suited to least-squares analysis. Poisson-based regression models of counts of offenses are preferable because they are built on assumptions about error distributions that are consistent with the nature of event counts. A simple elaboration transforms the Poisson model of offense counts to a model of per capita offense rates. To demonstrate the use and advantages of this method, this article presents analyses of juvenile arrest rates for robbery in 264 non metropolitan

counties in four states. The negative binomial variant of Poisson regression effectively resolved difficulties that arise in ordinary least-squares analyses.

Bennell and Canter (2002) use statistical models to test directly the police practice of utilizing modus operandi to link crimes to a common offender. Data from 86 solved commercial burglaries committed by 43 offenders are analyzed using logistic regression analysis to identify behavioral features that reliably distinguish between linked and unlinked crime pairs. Receiver operating characteristic analysis is then used to assign each behavioral feature an overall level of predictive accuracy. The results indicate that certain features, in particular the distances between burglary locations, lead to high levels of predictive accuracy. This study therefore reveals some of the important consistencies in commercial burglary behavior. These have theoretical value in helping to explain criminal activity. They also have practical value by providing the basis for a diagnostic tool that could be used in comparative case analysis.

In this article, the authors examine whether subjective perceptions of community safety are informed by the built environment. They posit that the built environment serves as a heuristic device, providing cues about likely levels of neighborhood crime, independent of the effects of neighborhood crime itself. Using data on 4,456 individuals nested within 100 census tracts, the authors estimate hierarchical logistic models of perceived community crime risk. They focus on the role of the neighborhood built environment in the form of aggregated perceptions of residential land use, while controlling for individual-level criminal opportunity, community-level social structural antecedents, and community-level objective crime. The findings indicate that the neighborhood-level presence of businesses and parks and playgrounds increases individual perceptions of community danger, but these effects disappear once neighborhood crime rates are controlled. The presence of schools has no effect on subjective interpretations of community crime, regardless of whether actual area crime is considered (Wilcox *et al.*, 2003).

Dinkes (2004) research is concerned with logistic multilevel modelling of public satisfaction with the London Metropolitan Police Service. Data from the 2000 Policing for London-Responding to Diversity survey, merged with The Indices of Deprivation 2000 from the Office of the Deputy Prime Minister, creates the complete data set. Approval ratings are an important indicator of the potential success of new police initiatives. High levels of satisfaction correlate to an increased likelihood of reporting crime, thus enabling the police to provide better services. Following previous research and theory, individual level variables are tested and selected for the final model. Ward identifiers and the index of deprivation comprise the level-two variables. Previous research identifies neighbourhood environments as an important predictor of satisfaction, but research has been slow to model or understand neighbourhood effects. Furthermore, previous models failed to model within group dependence, i.e. the likelihood that citizens in the same Ward will be more similar than individuals in different Wards. Consequently, simple models underestimated standard errors and yielded overly narrow confidence intervals. This paper seeks to avoid such fallacies by employing multilevel models to study satisfaction with the police. The findings of this research conclude that neighbourhood effects persist after controlling for area deprivation and possibly attenuate the effects of race.

Davis (2006) used a cross section of 636 Los Angeles neighborhoods in 1987 to examine relationships between neighborhood rates of property crime, levels of policing as measured by neighborhood arrest rates, jobs per square mile, and characteristics of neighborhood residents. Endogeneity between neighborhood crime and arrest rates is explored by estimating a regression model with and without control variables for neighborhood characteristics and fixed area effects. When comparing nearby neighborhoods with similar characteristics, crime is lower where arrest rates are higher, but when comparing neighborhoods across the city without controls, arrest rates are higher in higher crime, lower income minority areas.

# CHAPTER THREE

## 3. DATA AND METHODOLOGY

The major aim of this study is to determine whether a particular crime related variable has an effect on a categorical response variable of crime severity. The study design should take in to account the sample size  $n$ , to provide a good chance of detecting an effect of a given size.

### 3.1 Sample Size Determination

The three criteria to determine appropriate sample size are the level of precision, the level of confidence or risk, and the degree of variability in the attributes being measured. A critical component of sample size formulas is the estimation of variance in the primary variables of interest in the study. When estimating the variance of categorical (proportional) variable Krejcie and Morgan (1970) recommended that researchers should use 0.50 as an estimate of the population proportion. This proportion will result in the maximization of variance, which will also produce the maximum sample size. Therefore, the required sample size given as:

$$n_0 = \frac{(t)^2 * pq}{(d)^2} \dots\dots\dots (1)$$

where  $t$  is a value for the selected alpha level (The alpha level indicates the level of risk the researcher is willing to take that true Margin of error may exceed the acceptable margin of error) and  $(p)(q)$  is estimate of variance = 0.25 to produce maximum possible sample size. Where  $d$  is acceptable margin of error for proportion being estimated (error researcher is willing to take).

Therefore, for a population of size  $N$ , the required initial sample size is  $n_0$ . However, since this sample size  $n_0$  exceeds 5% of the population ( $N * 0.05 = a$ ), Cochran's (1977) correction formula should be used to calculate the final sample size. These calculations are as follows:

$$n_1 = \frac{n_0}{1 + \frac{n_0}{N}} \dots\dots\dots (2)$$

Where:  $N$  is population size,  $n_0$  is the required initial sample size according to Cochran's formula and  $n_1$  is required sample size if the initial sample is greater than 5% of population. Therefore, the minimum sample size employed should be  $n_1$ . For this study the sample size  $n_1$  is selected using Simple Random Sampling (SRS).

### 3.2 Study Area

#### *Climate and Geography*

Tigray region is one of the nine regions of Ethiopia. The region is located in the northern part of the country and bordered with Eritrea to the north, Sudan to the west, Afar region to the east and Amhara region to the south (FDREPCC, 2008)

Tigray region has an estimated total population of 4,314,456, having male 2,124,853 (49.2%) and female 2,189,603 (50.8%). The annual growth rate is 2.5% (FDREPCC, 2008). The region has 5 zones, 35 districts and 694 kebeles and the capital city of the region is Mekelle. The total area of the region is about 50,078.64 square kilometers and most of the population lives in rural area (80.5). Of all urban dwellers in Ethiopia (11,951,170 people), 842,723 (or 7.05%) are living in Tigray (CSA, 2007)

Tigray has 7 ethnic groups. The cultural beliefs and religion among various ethnic groups were different (FDREPCC, 2008). Tigray region has low income and almost half of the population lives below poverty line; the economy is based on subsistence farming like other regions of Ethiopia (FMOHE, 2005). Tigray has low literacy level of adult. Literacy level of Tigray is for men 67.5% and women 33.7 %. This shows that men have great chance to education than female which amounts to inequality (EDHS, 2005).

### **3.3 Data source**

Crimes are recorded by police on daily basis. This study is based on a secondary data obtained from Tigray regional police commission for a particular year 2009/10.

The collected data has two segments: crime scene (Scene Label) and crime label (Crime Label). Scene Label refers to the place where the crime is committed, whether it is urban or rural. On the other hand, crime label refers to the label or seriousness of crime such as serious, medium, or low. Each record in the data set was classified into one of the variable values of each of the two attributes.

The whole purpose of defining these classes was to investigate which section of the society (individual having some values for the given attributes) is being involved in crime. A crime could be committed in urban or rural area and there is a need for police officers to know the rate and nature of crimes that are occurring in the respective areas. In addition, it is also equally important to identify a section of the society that involve in different category of crime depending on their severity. Thus, to assist the police commission in deploying appropriate resources allocation and crime prevention methods, the study attempted to identify highly committed of crime category, individuals commit in each category of crime and area of crime committed.

#### **Variables Considered**

The most important and common factors that influence the incidence of crime are described as follows.

**Demographic Variables:** The demographic variables related to individual involved in crime are:

- Age: it is categorized as:
  - 12-30,
  - 31-50 and
  - 51 and above

- Sex: (Male or Female)
- Educational Background:
  - No education
  - Primary education
  - Junior Secondary education
  - Secondary education
  - Higher education
- Employment Status :
  - Government employee
  - Self employ
  - Private employee
  - Farm employee
  - Unemployed
- Marital Status :
  - Single
  - Married

**Location related variable:** This variable indicates the area where crime is committed.

- Rural
- Urban

### **Dependent categorical Variable**

The dependent variable is crime severity. This includes serious, medium, and low crime. Crime is a composite variable which includes serious crimes such as murder, corruption, crimes against public institutions and the like, medium crimes includes rape, vandalism, fraud, and crimes against individual personality and low crimes such as smuggling trade,

drug use and transmission and so on. There are forty three types of crime in the region and all those types are classified in to three categories based on the cost effect of crime in the society. The classification of crime is performed by regional law enforcement agencies. The categorization has taken in to account minimum number of years the offender is penalized according to different types of crimes.

### **3.4 MODEL**

In order to meet the objective set up on this study Ordinal logistic regression model and tests related are employed as a general methodology.

#### **Logistic Regression Model**

The logistic model, as a non-linear regression model, is a special case of generalized linear model (McCullagh and Nelder, 1989) where the assumptions of normality and constant variance of residuals are not satisfied. This model is a statistical technique for predicting probability of an event, given a set of predictor variables. The procedure is more sophisticated than the linear regression procedure.

Logistic regression is used to predict the probability of dependent variable on the basis of independent variables and to determine the effect size of the independent variables on the dependent; to rank the relative importance of independents; to assess interaction effects; and to understand the impact of covariate control variables. The impact of predictor variables is usually explained in terms of odds ratio and hence the name logistic regression, also called the log-odds function. This model applies maximum likelihood estimation after transforming the dependent into a logit variable (the natural log of the odds of the dependent occurring or not).

#### **Assumptions of Logistic Regression**

The validity of inferences drawn from modern statistical modeling techniques depends on the assumptions of the statistical model being satisfied. In order to valid the analysis the model should satisfy the following assumptions.

- i.* It does not need a linear relationship between the dependent and independent variables.
- ii.* The error terms need to be independent. Logistic regression requires each observation to be independent. That is, the data-points should not be from any dependent samples design, e.g., before-after measurements, or matched pairings. Also the model should have little or no multicollinearity. However, there is the option to include interaction effects of categorical variables in the analysis and the model. If multicollinearity is present centering the variables might fix, i.e. deducting the mean of each variable. If this does not lower the multicollinearity a factor analysis with orthogonally rotated factors should be done before the logistic regression is estimated.
- iii.* Logistic regression assumes linearity of independent variables and log odds; it requires that the independent variables are linearly related to the log odds. Otherwise the logistic regression underestimates the strength of the relationship and rejects the relationship easily, that is being not significant (not rejecting the null hypothesis) where it should be significant. A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the logistic regression model. Another approach would be to use Discriminant analysis, if the assumptions of homoscedasticity, multivariate normality, and no multicollinearity are met.
- iv.* Logistic regression requires quite large sample sizes. Because in the case of small sample size maximum likelihood estimates are less powerful than ordinary least squares (e.g., simple linear regression, multiple linear regression).

### **3.5. Ordinal logistic regression model**

Logistic regression model can be classified as multinomial, ordinal and binary. In this investigation Ordinal logistic regression model was used. The ordinal logistic regression procedure empowers one to select the predictive model for ordered dependent variables. It describes the relationship an ordered response variable and a set of explanatory variables. The explanatory variables may be continuous or discrete (or any type).

Ordinal response models have major importance in social sciences as well as demography and many social phenomena. The responses are discrete or qualitative rather than continuous or quantitative in nature. Many such analyses involve an outcome or dependent variable that is ordinal and in these studies the logistic regression model has become the statistical model of choice. The most popular model in ordinal logistic is the Proportional Odds model.

### **3.6 Proportional Odds (PO) Model**

Proportional Odds Model is used as a tool to model the ordinal nature of a dependent variable by defining the cumulative probabilities differently instead of considering the probability of an individual event. It considers the probability of that event and all events that are ordered before it. When response categories are ordered, logits can directly incorporate the ordering. The cumulative probabilities are the probability that the response  $Y$  falls in category  $i$  or below, for each possible  $i$  the  $i^{th}$  cumulative probability is  $pr(Y \leq i) = p_1 + p_2 + \dots + p_i$ .

The proportional odds model assumes that the cumulative logits can be represented as parallel linear functions of independent variables. That is, for each cumulative logit the parameters of the models are the same, except for the intercept. Consequently, according to the proportional odds assumption, odds ratio is the same for all categories of the response variable.

The PO model, however, has some appealing features. At first, it is invariant under several categories, as only the signs of the regression coefficients change. Secondly, it is invariant under collapsibility of the ordered categories, as the regression coefficients do not change when response categories are collapsed or the category definitions are changed. Thirdly, it produces the most easily interpretable regression coefficients, as  $\exp(\beta)$  is the homogenous odds ratio over all cut-off points summarizing the effects of the explanatory factor  $X$  on the response  $Y$  in one single frequently used measure. Due to these reasons, the PO model is by far the most used regression model for ordinal data.

Let  $Y$  takes categorical response variable with  $c$  ordered categories and assume  $pr(Y = 1)$  is  $p_1$ ,  $pr(Y = 2)$  is  $p_2$ , ...,  $pr(Y = i)$  is  $p_i$ ; for  $i = 1, \dots, c$ . Cumulative probability reflect the ordering, with  $pr(Y \leq 1) \leq pr(Y \leq 2) \leq \dots \leq pr(Y \leq c) = 1$  and let the cumulative probability of the first  $c - 1$  of  $Y$  is  $pr(Y \leq i) = \pi_i, i = 1, \dots, c - 1$ .

Then the odds of the first  $c - 1$  cumulative probabilities are

$$odds(pr(Y \leq i)) = \frac{pr(Y \leq i)}{1 - pr(Y \leq i)} = \left[ \frac{\pi_i}{1 - \pi_i} \right] \quad i = 1, \dots, c - 1 \quad \dots \dots \dots (3)$$

The proportional Odds model models the log odds of the first  $c - 1$  cumulative probabilities as:

$$\begin{aligned} \log it[pr(Y \leq i)] &= \log \left[ \frac{pr(Y \leq i)}{1 - pr(Y \leq i)} \right] = \log \left[ \frac{\pi_i}{1 - \pi_i} \right] \\ &= \log \left[ \frac{\pi_i}{1 - \pi_i} \right] \dots \dots \dots (4) \end{aligned}$$

And the relationship between the cumulative logits of  $Y$  is:

$$\log \left[ \frac{\pi_i}{1 - \pi_i} \right] = \log \left[ \frac{\pi_i}{\pi_{i+1} + \dots + \pi_c} \right]; i = 1, \dots, c - 1.$$

Consider a collection of  $P$  explanatory variables denoted by the vector  $X' = (X_1, X_2, \dots, X_p)$ . The relationship between the predictor and response variables is not a linear function in logistic regression; instead, the logistic regression function is used, which is the logit transformation of  $\pi$ .

$$\pi_i = \frac{\exp(\alpha_i + \beta_1 X_1 + \dots + \beta_p X_p)}{1 + \exp(\alpha_i + \beta_1 X_1 + \dots + \beta_p X_p)} \dots \dots \dots (5)$$

Then the logit or log-odds of having  $pr(Y \leq i) = \pi_i$  is modeled as a linear function of the explanatory variables as:

$$\log \left[ \frac{pr(Y \leq i)}{1 - Pr(Y \leq i)} \right] = \log \left[ \frac{\pi_i}{1 - \pi_i} \right] = \alpha_i + \beta_1 X_1 + \dots + \beta_p X_p.$$

Equivalent with

$$\log \left[ \frac{\pi_i}{1 - \pi_i} \right] = \alpha_i + \sum_{j=1}^p \beta_j X_j; 0 \leq \pi_i \leq 1; \text{ Therefore}$$

$$\text{logit}[pr(Y \leq i)] = \alpha_i + \sum_{j=1}^p \beta_j X_j \quad i = 1, \dots, c-1 \quad \text{and} \quad j = 1, \dots, p \dots \dots (6)$$

The model assumes a linear relationship for each logit and parallel regression lines. Equation (6) is called proportional odds model and it estimates simultaneously multiple equations of cumulative probability. An equation is solved for each category of the dependent variable except the last one.

In this model each logit has its own  $\alpha_i$  term called the threshold value and their values do not depend on the values of the independent variable for a particular case. Logistic regression coefficients are indicates the direction and strength of the relationship between independent variable and the log odds of dependent variable. However, these logistic regression coefficients are a little bit more complicated to intuitively gauge, as they present the influence of a unit change in the independent variable on the log odds of the dependent variable. The influence determines the rate of increase or decrease in the log odds of dependent variable. This means that the effect of the independent variable is the same for different logit functions, that's also the reason why the model is called the proportional odds model.

### **Testing of Parallel Lines**

The assumption that all logit surfaces are parallel must be tested. Test of parallel lines helps to determine whether it is reasonable to assume that the values of the location

parameters are constant across categories of the response. The test of parallelism contains:  $-2 \log$ -likelihood for the constrained model, the model that assumes the planes or surfaces are parallel and  $-2 \log$ -likelihood for the General model, the model that assumes planes or surfaces are separated.

The chi-square statistic is the log-likelihood difference between the two models. If the lines or planes are parallel, the observed significance level for the change should be large, since the general model doesn't improve the fit very much and the parallel model is adequate. If there is evidence to reject the null hypothesis, it is possible that the link function selected is incorrect or that the relationships between the independent variables and logits are not the same for all logits.

### **3.7 Partial Proportional Odds Model**

As the proportional odds assumption is difficult to achieve in practice, the PPOM may be used as an alternative. This model allows some covariables with the proportional odds assumption to be modeled, but for those variables in which this assumption is not satisfied it is increased by a coefficient ( $\gamma$ ), which is the effect associated with each  $i^{th}$  cumulative logit, adjusted by the other covariables. The general form of the model is the same as the PO model, but now the coefficients are associated with each category of the response variable.

Partial proportional odds model can be classified as PPOM-UR and the restricted one. The unrestricted partial proportional odds model is used when proportional chances assumption is not valid and the coefficients are associated with each category of the response variable (in the case of both parallel and linear assumption are not fulfilled). The model has the form:

$$\lambda_i = \ln \left\{ \frac{pr(Y = 1 / \vec{X}) + \dots + pr(Y = i / \vec{X})}{pr(Y = (i+1) / \vec{X}) + \dots + pr(Y = k / \vec{X})} \right\} = \ln \left\{ \frac{\sum_{1 \rightarrow}^i pr(Y = i / \vec{X})}{\sum_{i+1 \rightarrow}^k pr(Y = i / \vec{X})} \right\}$$

$$\lambda_i = \alpha_i + \{(\beta_1 + \gamma_{i1})X_1 + \dots + (\beta_q + \gamma_{iq})X_q + (\beta_{q+1}X_{q+1}) + \dots + (\beta_p X_p)\}, i = 1, \dots, k-1 \dots (7)$$

It is normally expected that there will be a type of linear trend between each OR of the specific cut-off points and the response variable. If there is then a set of restrictions  $\gamma_{kl}$  may be included in the model to clarify this linearity. When these restrictions are included this model is called the restricted partial proportional odds model. The  $\tau_i$  parameters are fixed scale parameters which take the form of restrictions allocated to the parameters. In this case for a given covariable  $X_m$ ,  $\alpha_m$  does not depend on the cut-off points, but is multiplied by  $\tau_i$  for each  $i^{th}$  logit. The model becomes:

$$\lambda_i = \ln \left\{ \frac{pr(Y = 1 / \vec{X}) + \dots + pr(Y = i / \vec{X})}{pr(Y = (i+1) / \vec{X}) + \dots + pr(Y = k / \vec{X})} \right\} = \ln \left\{ \frac{\sum_{1 \rightarrow}^i pr(Y = i / \vec{X})}{\sum_{i+1 \rightarrow}^k pr(Y = i / \vec{X})} \right\}$$

$$\lambda_i = \alpha_i + \{\tau_i ((\beta_1 + \gamma_{i1})X_1 + \dots + (\beta_q + \gamma_{iq})X_q + (\beta_{q+1}X_{q+1}) + \dots + (\beta_p X_p))\}, i = 1, \dots, k-1 \dots (8)$$

### 3.8 Odds Ratio

The odds ratio is a value which measures the strength of effect of each independent variable in the model on the log odds of the dependent variable.

The odds of some event happening is defined as the ratio of the number of occurrences to the number of non occurrences. That is, the odds of the event  $E$  is given by:

$$odds(E) = \frac{pr(E)}{pr(not E)} = \frac{pr(E)}{1 - pr(E)} \dots\dots\dots (9)$$

The odds of the response are multiplied by  $e^\beta$  for every unit increment of  $x$ . That is, the odds at level  $x + 1$  equal the odds at  $x$  multiplied by  $e^\beta$  and odds less than one indicate the occurrence is less likely than non occurrence.

### 3.9 Model Selection

Methods such as forward, backward, and stepwise selection are available, but, in logistic as in other regression methods are not to be recommended. They give incorrect estimates of the standard errors and p-values, can discard variables that are important to be included in the model (Harrell, 2001). It is much better to compare models based on their results, reasonableness, and fit as measured, by the Akaike Information Criterion. Other criteria besides significance tests can help select a good model in terms of estimating quantities of interest. The best known is the Akaike information criterion. It judges a model by how close its fitted values tend to be to the true values, in terms of a certain expected value. Even though a simple model is farther from the true model than is a more complex model, it may be preferred because it tends to provide better estimates of certain characteristics of the true model, such as cell probabilities. Thus, the optimal model is the one that tends to have fit closest to reality. Given a sample, Akaike showed that this criterion selects the model that minimizes

$$AIC = -2(\text{maximized log likelihood} - \text{number of parameters in model}).$$

This penalizes a model for having many parameters. With models for categorical  $Y$ , this ordering is equivalent to one based on an adjustment of the deviance,  $(G^2 - 2(d.f.))$ , by twice its residual degree of freedom (Agresti, 2002).

### Test of Overall Model Fit

For the selected model before proceeding to examine the individual coefficients, we should look at an overall test of the null hypothesis that the location coefficients for all of the variables in the model are 0.

It can base this on the change in  $-2 \log$ -likelihood when the variables are added to a model that contains only the intercept. The change in likelihood function has a chi-square distribution even when there are cells with small observed and predicted counts. This value provides a measure of how well the model fits the data. The log likelihood statistic is analogous to the error sum of squares in multiple linear regressions. As such it is an indicator of how much unexplained information remains after fitting the model. The larger the value of the log likelihood the more unexplained observations there are and a poorly fitting model. Therefore, a good model means a small value for  $-2LL$ . If a model fits perfectly, the likelihood is 1, and  $-2 \times \log 1=0$ .

### 3.10 Goodness-of-Fit Measures

A good-fitting model has several benefits. The structural form of the model describes the patterns of association and interaction. The sizes of the model parameters determine the strength and importance of the effects. Inferences about the parameters evaluate which explanatory variables affect the response variable  $Y$ , while controlling effects of possible confounding variables. Finally, the model's predicted values smooth the data and provide improved estimates of the mean of  $Y$  at possible explanatory variable values.

For logistic regression, the model coefficients are estimated by the maximum likelihood method and the likelihood equations are non-linear explicit function of unknown parameters. The ordinal logistic regression model is fitted to the observed responses using the maximum likelihood approach. In general, the method of maximum likelihood produces values of the unknown parameters that best match the predicted and observed probability values. Therefore, it usually used a very effective and well known Fisher scoring algorithm to obtain ML estimates.

A model for logit  $pr(Y \leq i)$  alone is ordinary logit model for a binary response in which categories 1 to  $i$  form one outcome and categories  $i+1$  to  $c$  form a second outcome. This shows that  $c$  categories of response collapsed in to binary out come. Again let  $(Y_{j1}, \dots, Y_{jc})$  be binary indicators of the response for subject  $j$ .

The likelihood function  $L$  is viewed as a function of  $\beta$  and  $\alpha_i$  parameters. The parameters are estimated by maximizing the likelihood, or more usually, by maximizing the logarithm of the likelihood. The likelihood function is given by the equation:

$$L = \prod_{j=1}^n \left[ \prod_{i=1}^c \pi_i(X_j)^{Y_{ij}} \right] = \prod_{j=1}^n \left[ \prod_{i=1}^c (p(Y \leq i / X_j) - p(Y \leq i-1 / X_j))^{Y_{ij}} \right]$$

$$= \prod_{j=1}^n \left[ \prod_{i=1}^c \left( \frac{\exp(\alpha_i + \beta' X_j)}{1 + \exp(\alpha_i + \beta' X_j)} - \frac{\exp(\alpha_{i-1} + \beta' X_j)}{1 + \exp(\alpha_{i-1} + \beta' X_j)} \right)^{Y_{ij}} \right]$$

$$l(\beta^*) = \prod_{j=1}^n \left[ \pi_1(X_j)^{Y_{1j}} \pi_2(X_j)^{Y_{2j}} \times \dots \times \pi_c(X_j)^{Y_{cj}} \right]$$

Here  $\beta^*$  use somewhat imprecisely to denote both the slope coefficients and intercept coefficients. It follows that the log-likelihood function is:

$$L(\beta^*) = \sum_{j=1}^n Y_{1j} \ln[\pi_1(X_j)] + Y_{2j} \ln[\pi_2(X_j)] + \dots + Y_{cj} \ln[\pi_c(X_j)] \dots \dots \dots (10)$$

The maximum possible value of the likelihood for a given data set occurs if the model fits the data exactly. This occurs if observed counts are close with predicted. The difference between the log-likelihood functions for two models is a measure of how much one model improves the fit over the other. A special case of this was defined as the deviance. The deviance is defined as minus twice the log of the ratio of the likelihood for a model to the maximum likelihood. Deviance for model comparison is:

$$D = -2 \log \left[ \frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}} \right], \text{ This simplify to}$$

$$D = -2[\log(\text{likelihood of current model}) - \log(\text{likelihood of saturated model})]$$

$$= \{(-2 \log \text{likelihood of current model}) - (-2 \log \text{likelihood of saturated model})\}$$

The deviance can be shown that the likelihood of this saturated model is equal to 1 yielding a log-likelihood equal to 0. For a sample of  $n$  independent observations, the deviance for a model with  $p$  degrees of freedom (that is,  $p$  parameters estimated, including the threshold or constant) has  $(n - p)$  degrees of freedom.

Since the deviance is effectively -2 times the log of the likelihood ratio, it has an asymptotic distribution that is chi-squared with degrees of freedom equal to  $(n - p)$ .

This deviance is also used to construct a goodness-of-fit test for the model. The goodness of fit statistics for ordinal logistic regression has a form:

$$D = 2 \sum \sum O_{ij} \log \left( \frac{O_{ij}}{E_{ij}} \right)$$

Likewise, the Pearson chi-square statistic also compares the model fit to the actual data, defined by:

$$\chi^2 = \sum \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}; \quad E_{ij} \text{ is the expected value for the } i^{\text{th}} \text{ observation.}$$

Both goodness-of-fit statistics should be used only for models that have reasonably large expected values in each cell. If the model fits well, the observed and expected cell counts are similar, the value of each statistic is small, and the observed significance level is large. As usual large  $\chi^2$  and  $D$  value provide the evidence of lack of fit. When the fit is poor, residuals and other diagnostic measure describes the influence of individual observation on the model fit and highlight reason for the inadequacy.

### 3.11 Model Adequacy

One important topic in logistic regression is regression diagnostics. In logistic regression, some of the usual diagnostic statistics are the residuals and measures of influence.

## Model Evaluation- Residuals

Residuals are the basic building blocks for logistic regression diagnostics. They can be useful for identifying potential outliers (observations not well fit by the model) or misspecification models and another use for residuals is in checking normality. For log-linear models this can be thought of checking how well the asymptotic theory holds.

The residuals for logistic regression model are typically defined as the difference between observed response, and the estimated probability of the response, conditional on the covariates. This simplifies with categorical predictors of  $Y$ , it is useful to form residuals to compare observed and fitted counts. Let  $Y_i$  denote the binomial variate for  $n_i$  trials at setting  $i$  of the explanatory variables  $i = 1, \dots, N$ . Let  $\hat{\pi}_i$  denote the model estimate of  $p(Y = 1)$ . Then  $n_i \hat{\pi}_i$  is the fitted number of successes. However, in logistic regression we have binomial errors and, as a result, the error variance is a function of the conditional mean. For a GLM with binomial random component, the Pearson residual for this fit is:

$$e_i = \frac{Y_i - n_i \hat{\pi}_i}{\left[ \widehat{\text{var}}(Y_i) \right]^{1/2}} = \frac{Y_i - n_i \hat{\pi}_i}{\sqrt{\left( n_i \hat{\pi}_i (1 - \hat{\pi}_i) \right)}}$$

Which standardizes by dividing the difference by the estimated binomial standard deviation of  $Y_i$ . These residuals relate to the Pearson goodness-of fit statistics by

$$\chi^2 = \sum_{i=1}^N e_i^2$$

Each squared Pearson residual is a component of  $\chi^2$ . Pearson residual values fluctuate around zero, following approximately a normal distribution when  $n_i$  is large. When the model holds these residuals are less variable than standard normal (that is  $e_i$  has an approximate  $N(0,1)$ ), however, because the numerator must use the fitted value  $n \hat{\pi}$  rather than the true mean  $n \pi$ .

Since the sample data determine the fitted value,  $Y_i - n_i \hat{\pi}_i$  tends to be smaller than  $Y_i - n_i \pi_i$ . The Pearson residual divided by its estimated standard error is called an adjusted residual. Adjusted residual larger than about 2 in absolute value are worthy of attention, though one expect some values of this size by chance alone when the number of categories is large. Adjusted residuals are preferable to Pearson residuals (Agresti, 1996).

Deviance residual is another type of residual. It measures the disagreement between the maxima of the observed and the fitted log likelihood functions. The deviance residual is useful for determining if individual points are not well fit by the model. The deviance residual for the  $i^{th}$  observation is the signed square root of the contribution of the  $i^{th}$  case to the sum for the model deviance, for the  $i^{th}$  observation, and is given by

$$D_i = \pm \left\{ -2 \left[ Y_i \log \hat{\pi}_i + (1 - Y_i) \log (1 - \hat{\pi}_i) \right] \right\}^{1/2},$$

Where the sign is positive when  $Y_i \geq \hat{\pi}_i$  and negative otherwise. An observation with a residual that is far from 0 (that is greater than two in either direction) is poorly fit by the model.

### **Influence Measures**

It measures the effect that deleting an observation has on the regression parameters or the goodness-of-fit statistics. An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness. It may be informative to report the fit of the model after deleting one or two observations, if the fit with them seems misleading.

Leverages are the diagonal elements of the logistic equivalent of the hat matrix in general linear regression. The  $i^{th}$  diagonal element of the logistic equivalent of the hat matrix is calculated as:

$$h_i = n_i \pi(\hat{X}_i) \left[ 1 - \hat{\pi}(X_i)(1, X'_j)(X'VX)^{-1}(1, X'_j)' \right], \quad \text{Where } \hat{V} = \text{diag} \left\{ \hat{\pi}_i(1 - \hat{\pi}_i) \right\} \text{ and}$$

$\hat{\pi}_i$  is the expected proportional response with  $n_i$  number of trials of the  $i^{\text{th}}$  covariate pattern. An observation with an extreme value on a predictor variable is called a point with high leverage. Leverage is a measure of how far (proportional distances of) an independent variable deviates from its mean. These leverage points can have an effect on the estimate of regression coefficients and its value measures the influence of a point on the fit of the model. The centered leverage ranges from 0 (no influence on the fit) to  $(N-1)/N$ , and a leverage value greater than 2 or 3 times of average leverage is considered as large.

The logistic regression analog of Cook's influence statistic is a measure of how much the residual of all cases would change if a particular case were excluded from the calculation of the regression coefficients. Cook's distance is a direct influence measure relative to the fitted regression coefficients and observation with higher Cook's distance is the more influential point. For logistic regression the Cook's distance has the form:

$$CD_i = \frac{Z_i^2 h_i}{1 - h_i}, \quad \text{Where } Z_i \text{ is Standardized Residual.}$$

A large Cook's distance indicates that excluding a case from computation of the regression statistics changes the coefficients substantially. The lowest value of Cook's distance can assume is zero but for logistic regression, a case is identified as influential if its Cook's distance is greater than one (Hosmer and Lemeshow, 2000).

# CHAPTER FOUR

## 4. RESULT AND DISCUSSION

### 4.1 Descriptive statistics

This study used crime data which is collected from regional police commission of Tigriy region. Out of the total population 40,744 the sample size 3,620 is determined by employing 2% margin of error, 1% risk level and 0.5 population proportion. Of the total the subjects with complete criminal information is 2,753. Table 1, below, shows that from the sampled crime data 40.6% are serious crime, 32.7% are medium crime and 26.7% are low crime.

Table1: distribution of records based on crime level (crime scene)

Crime status	Frequency	percent
Serious	1118	40.6
Medium	901	32.7
Low	734	26.7
Total	2753	100

#### A. Criminal related variables

##### *Criminal's age*

The criminal's age has three categories. Among these categories, criminals with the age group 12-30 are responsible for large number of crime (60.2%) and criminals with age 51 and above have the smallest share (Table 14).

##### *Criminal's Educational Background*

Among the five categories of educational background of criminals, 46.6% of crime is committed by non-educated criminals. However, those criminals who have primary and junior education level are almost similar result and criminals who have higher education

level have the lowest share 4.5%. These results show that, crime is highly committed by those no educated criminals (Table 14).

#### *Criminal's Employment status*

The result of criminal employment status indicates unemployed criminals take the major share (30%) and those criminals who are farm employees take the second share (24.8%). However, criminals who are government and private employees have relatively similar result. Based on the result (Table 13 Annex 3) uneducated and unemployed individuals who commit crime are 27.4%, but those secondary and higher education as well as unemployed are 52.1%. This results show that, crime incidence is higher among educated and unemployed individuals than those uneducated and unemployed.

#### *Criminal's Marital Status*

This variable has two categories: single and married. The descriptive statistics indicates 54.9% of criminals belong to a single member household and the rest are married.

### **B. Location related Variables**

#### *Place of crime committed*

Among the two categories of places (rural and urban), the occurrence of crime is highest for rural area (which is 60.6%).

### **4.2 Logistic Regression Analysis**

This section focused on regression analysis undertaken to test the relative predictive power of socio-demographic and environmental covariates with severity of crime.

In this study ordinal logistic regression is selected for analyzing the crime data using the explanatory variables associated with the dependent variable. Accordingly, sex (SC), Age (AC), Educational Background (CED), Employment Status (CEM), Marital Status (CMS), and Area of crime committed (ACC) are included in the model.

To find correct estimates of standard errors and p-values it is necessary to choose better model. The main effect model has a total of  $n$  parameter which is equal to 21 and the interaction model has a total of  $m$  parameter equals to 166 including the two intercepts. To select the model, here, we consider deviance comparison. The deviance is calculated between main effects and interaction model.

$$D = \{(-2 \log \text{likelihood for main effects model}) - (-2 \log \text{likelihood for main effect and interaction model})\}$$

with degree freedom  $m - n = 166 - 21 = 145$ .

Form the output (A.1.1 and A.1.3) the deviance can be computed as:

$D = (5744.478 - 5819.587) = 75.109$  that follows  $\chi^2_{(145)}$  and non- significant at 5% level of significant with p- value 0.999. Thus, it can be conclude that there is no statistical benefit that can be gained by considering the more complex model, which is the interaction model.

In addition analyzing the data using SAS, the following results were obtained for full fledged model.

Table 2. Comparison of ordinal logistic regression model based on AIC criterion (full model)

Model	AIC
Null model	5972.303
Main effect model	5849.587
Interaction model	5904.478

The result of Akaike's Information Criteria (AIC) suggests that the main effect model is reasonable since a model with low AIC is preferred model. Therefore, a cumulative consideration of the above tests confirmed that the main effect model is appropriate.

The model takes Y values: 0(serious crime), 1(medium crime), and 2(low crime) and assume  $\pi_0 = pr(Y \leq 0)$ ,  $\pi_1 = Pr(Y \leq 1)$ ,  $\pi_2 = Pr(Y \leq 2)$ . The fitted proportional odds model of crime severity is given as follows:

$$\log\left[\frac{\pi_0}{(1-\pi_0)}\right] = \alpha_0 + \beta_{1h}SC_h + \sum_{i=1}^2 \beta_{2i}CA_i + \sum_{j=1}^4 \beta_{3j}CED_j + \sum_{k=0}^3 \beta_{4k}CEM_k + \beta_{5l}CMS_l + \beta_{6m}ACC_m$$

$$\log\left[\frac{(\pi_0 + \pi_1)}{1-(\pi_0 + \pi_1)}\right] = \alpha_1 + \beta_{1h}SC_h + \sum_{i=1}^2 \beta_{2i}CA_i + \sum_{j=1}^4 \beta_{3j}CED_j + \sum_{k=0}^3 \beta_{4k}CEM_k + \beta_{5l}CMS_l + \beta_{6m}ACC_m$$

where  $\left[\frac{\pi_0}{(1-\pi_0)}\right]$  and  $\left[\frac{(\pi_0 + \pi_1)}{1-(\pi_0 + \pi_1)}\right]$  are the log odds for respective cumulative logit model.

$\alpha_0$  and  $\alpha_1$  are threshold value for each model, respectively.

$SC_h$  represents female criminals for  $h = 1$ .

$CA_1$  and  $CA_2$  represent criminals with middle age group ( age between 31 and 50) and older age group (age between 51 and above).

$CED_j$  for  $j = 1, \dots, 4$  represent criminals education level where ( $j=1$  is primary education,  $j=2$  is junior secondary education,  $j=3$  is secondary education,  $j=4$  is higher education).

$CEM_k$  for  $k = 0, 1, 2, 3$  represent criminals employment status where ( $k=0$  is government employee,  $k=1$  is self employee,  $k=2$  is private employee,  $k=3$  is farm employee).

$CMS_l$  represent criminals marital status where  $l = 1$ , is married criminals.

$ACC_m$  represent area of crime committed where  $m = 1$ , is urban area.

The PROC Logistic procedure of SAS is used to generate coefficients of the estimated model. The complete output is presented in Annex1.1. From the results of SAS output, the two regression equation consisting of the above variables are given by (Table 4):

$$\begin{aligned}
\log\left[\frac{\pi_0}{(1-\pi_0)}\right] &= -0.27 + 0.20SC_1 - 0.12AC_1 - 0.86AC_2 - 0.08CED_1 - 0.10CED_2 - \\
&0.13CED_3 - 0.70CED_4 + 0.69CEM_0 + 0.63CEM_1 + 0.61CEM_2 + 0.48CEM_3 - 0.60CMS_1 \\
&- 0.23ACC_1 \\
\log\left[\frac{(\pi_0 + \pi_1)}{1 - (\pi_0 + \pi_1)}\right] &= 1.18 + 0.20SC_1 - 0.12AC_1 - 0.86AC_2 - 0.08CED_1 - 0.10CED_2 - \\
&0.13CED_3 - 0.70CED_4 + 0.69CEM_0 + 0.63CEM_1 + 0.61CEM_2 + 0.48CEM_3 - 0.60CMS_1 \\
&- 0.23ACC_1
\end{aligned}
\right.
\text{Model-1}$$

The result in (Annex 1.1) shows that sex of criminals is the only non-significant variable, where there is no significant difference between male and female. Since sex has non-significant effect, another model is fitted by excluding this variable. The new model is given as follow:

$$\begin{aligned}
\log\left[\frac{\pi_0}{(1-\pi_0)}\right] &= -0.25 - 0.12AC_1 - 0.85AC_2 - 0.09CED_1 - 0.11CED_2 - \\
&0.14CED_3 - 0.70CED_4 + 0.68CEM_0 + 0.63CEM_1 + 0.61CEM_2 + 0.48CEM_3 - 0.59CMS_1 \\
&- 0.24ACC_1 \\
\log\left[\frac{(\pi_0 + \pi_1)}{1 - (\pi_0 + \pi_1)}\right] &= 1.20 - 0.12AC_1 - 0.85AC_2 - 0.09CED_1 - 0.11CED_2 - \\
&0.14CED_3 - 0.70CED_4 + 0.68CEM_0 + 0.63CEM_1 + 0.61CEM_2 + 0.48CEM_3 - 0.59CMS_1 \\
&- 0.24ACC_1
\end{aligned}
\right.
\text{Model-2-}$$

### Goodness of Fit

It is useful to be able to judge whether the model fit to the data. A useful quantity in judging goodness of fit is the deviance. The saturated model has a total of  $n$  parameters. The regression model with fewer variables has  $K$  independent variables and 2 intercept. However, if the reduced model (model-2) is good, it should fit almost as well as the saturated one.

In case of the two fitted models, the full model (model-1) has 19 dummy variables with 2 intercepts which results  $n = 19 + 2 = 21$  parameters. The reduced model (model-2) has 17 dummy variables and 2 intercepts which results  $p = 17 + 2 = 19$  parameters.

To test the goodness of fit of model-2, we can use the deviance to compare its goodness of fit with model-1.

$Deviance = \{(-2 \log \text{likelihood model} - 2) - (-2 \log \text{likelihood model} - 1)\}$  With degrees of freedom equals to  $n - p = 21 - 19 = 2$ .

From the SAS output (Annex 1.1 and 1.2), the value of deviance is equal to  $5821.772 - 5819.587 = 2.185$  that follow  $X^2_{(2)}$  and non significant at 5% level of significance with p-value 0.335. Thus, it can be concluded that there is no statistical advantage that can be gained by considering the more complex model (model-1). Therefore, model-2 is selected as the better regression model.

Further, significance effect of the variables in model-2 was checked by fitting the Intercept-Only model. The deviance is  $5821.772 - 5968.303 = 146.531$ , it follows  $X^2_{(17)}$  and significant at 5% level of significance with  $p$ -value much less than 0.0001. Therefore, we can conclude that variables in model-2 have significant effect and the final model is a reduced model (model-2). Similar with the full model procedure the results from (A.1.2 and A.1.3 of the Annex) for the reduced model of which the deviance value equal to 61.312 with 111 d.f. is insignificant at 5% level and p-value of 0.99, and AIC shows that there is no evidence of a statistically significant interaction.

From main effect model of regression analysis, the proportional odds assumption is justified. The result of Score test with p-value (0.2647, Annex 1.2) shows that the observed significance value is large and the chi-square statistic is not significant at 5% level of significance. Consequently we do not reject the null-hypothesis, that planes or surfaces are parallel. This implies there is homogeneity of odds ratio for the two compared categories of response variable. We can show this graphically by taking age for illustration as follows and other graphs are shown in (Annex 4).

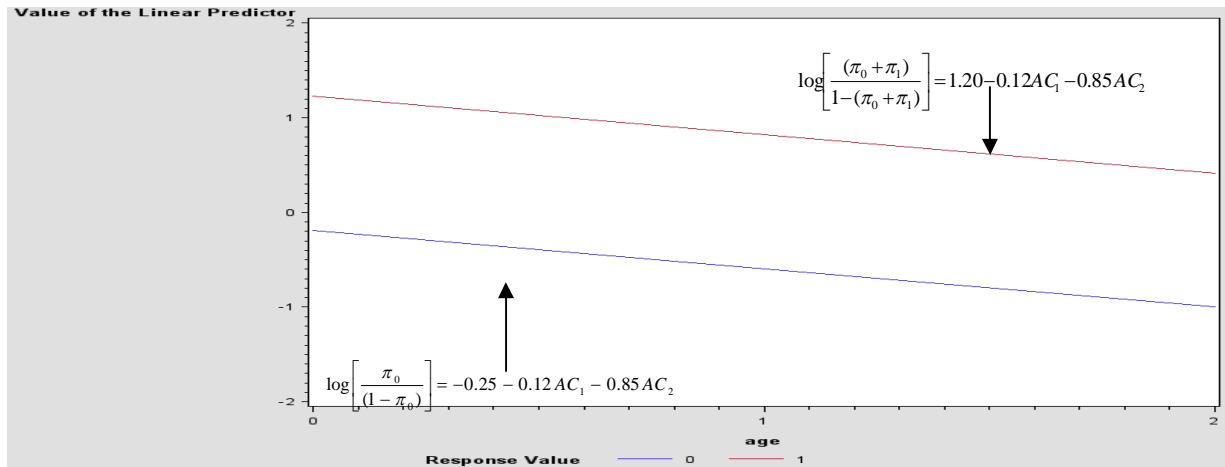


Figure 3: polts of cumulative logits of crime severity represented as Parallel linear function of variable Age.

The figure shows that logits of crime severity (i.e., logits of serious, and logits of serious or medium crime) are represented as parallel linear function of predictor (age). This indicates for each cumulative logit, the (age) parameter of the models is the same except for the intercept. Consequently, the odds ratio of the two categories (serious, and serious or medium crime) is homogenous since odds ratio has a value  $e^\beta$ . Therefore, it can be concluded that there is considerable evidence that lead to accepting the assumption of proportional odds model.

The goal of this study is to predict correctly the outcome of crime severity using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables which are useful in predicting the response variable. To satisfy this it is useful to check the adequacy of the selected model.

When the cumulative logit model fit well, it also fits well with similar effects for any collapsing of the response categories. Since diagnostic for ordinal and multinomial model is very difficult, one way to examine model adequacy is to check each of the binomial models separately (Hosmer and Lemshow, 2000). Therefore, diagnostic is performed

using binary outcome by collapsing medium and low crime in to moderate crime. Finally, the outcome variable becomes binary with serious and moderate crime.

Model diagnostic is performed for both outliers and influential points. To observe the model adequacy, one way of looking at them is to graph them against predicted probabilities.

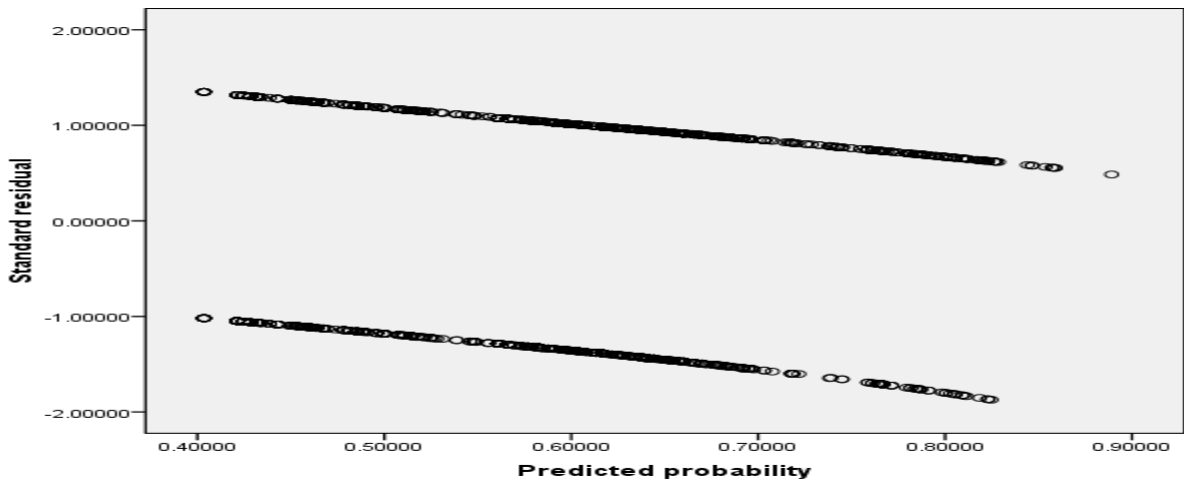


Figure 4: Plots of Standard residual by predicted probability.

Figure 4 shows that the density of points are higher on the upper curve for fitted values greater than 0.6, with residuals approaching 0. Similarly, the density of points on the lower curve is higher at fitted values less than 0.5, again with residuals near 0.

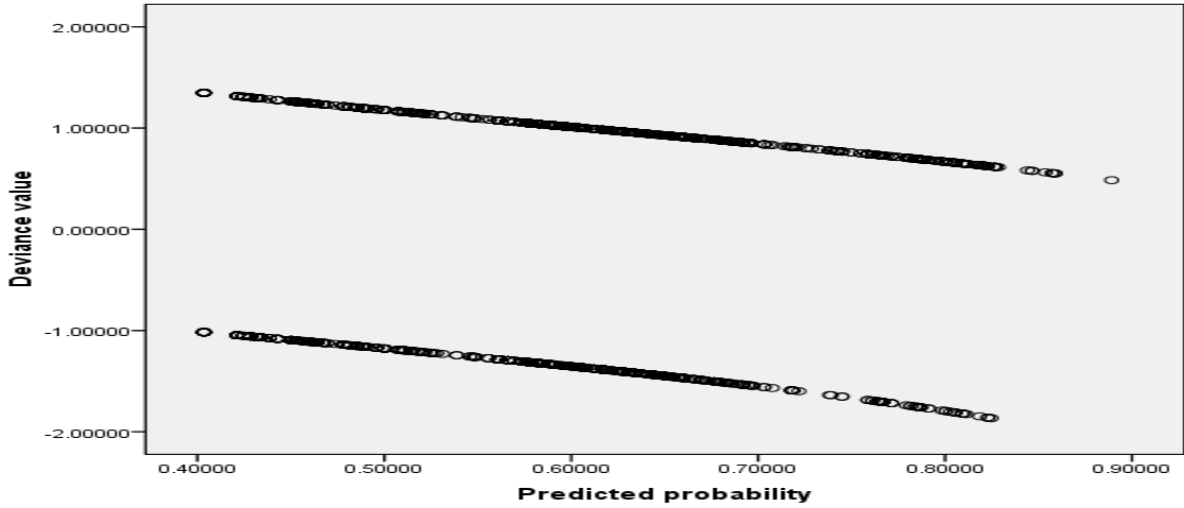


Figure 5: Plots of Deviance residual by predicted probability.

Figure 5 also confirms that points has similar pattern like the standard residual plots. In both cases none of the observation has standard and deviance residuals larger than 2 in absolute. As a result no observation can take worthy attention. This indicates the values for the independent variable are not in an extreme region. In addition observed responses for those points have similar form with the predicted probability of the response. Therefore individual points are well fitted by the model.

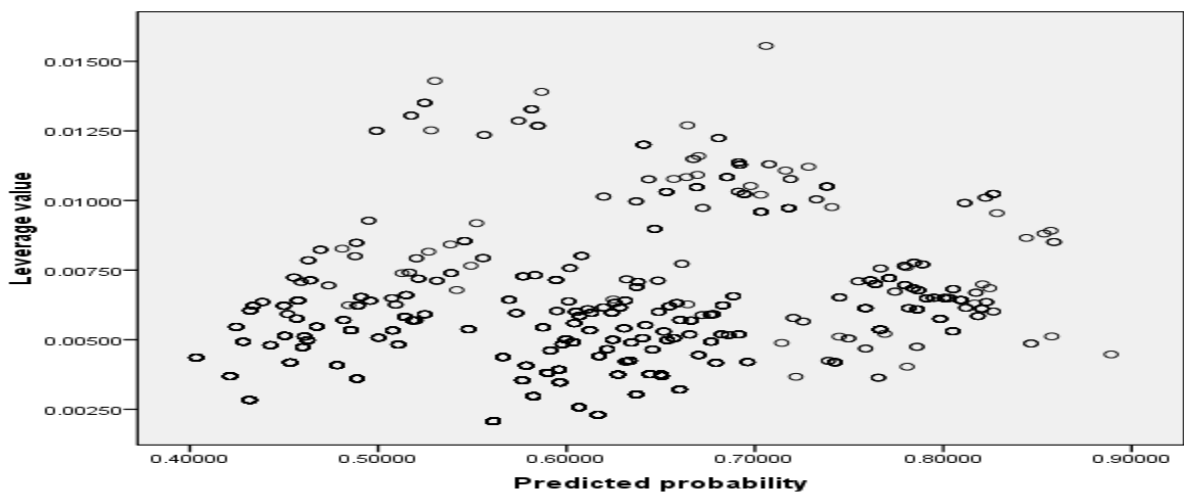


Figure 6: Plots of leverage value by predicted probability.

Plot of leverage against predicted probability is presented in Figure 6 and this identifies observations that have a strong influence on the estimated regression coefficients. The computed value of center of leverage is between 0 and 0.9996. The above plot shows that no point has large leverage value than two or three fold of 0.4998, which is average leverage. This indicates no observation is highly deviate from its mean and there is no observation with an extreme value on a predictor variable. In other words, the points have no undue influence on the parameter estimates of the fitted model. Therefore, the parameters are estimated properly and points are well fitted by the model.

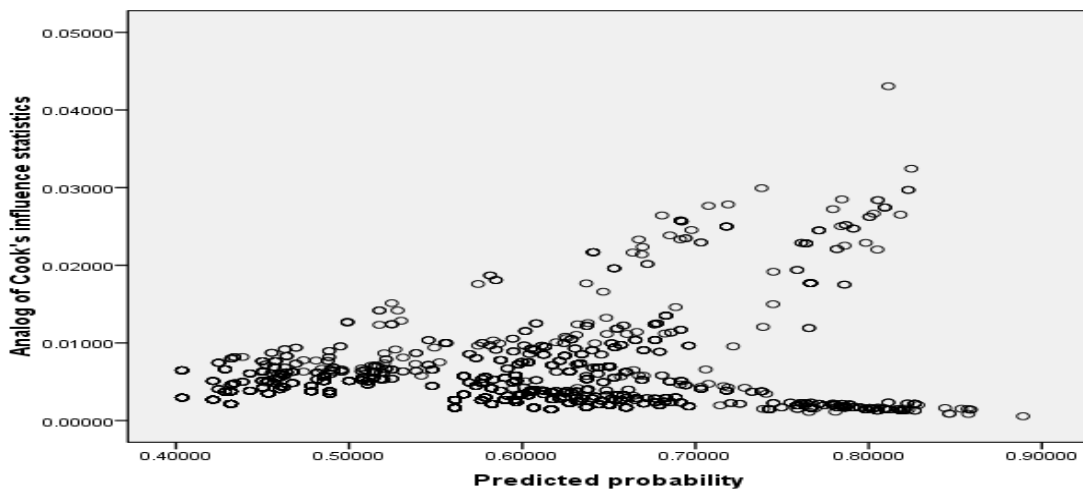


Figure 7: plots of cook's influence by predicted probability.

When looking at the plot of points in Figure 7 relatively few predicted values were found to be far from most of the other predicted values. Even if they are far away they may not be considered as real influential because no observation has values of Cook's distance greater than one. Therefore there is no observation with real influence and points are well fitted by the model.

Moreover, the overall model fit evaluates the contribution of each effect to the model. The results of Likelihood ratio, Score and Wald test for model goodness of fit displayed in table 3, suggests that model is well fitted to the data.

Table 3: The Likelihood Ratio, Score and Wald tests for overall measures of goodness of fit of the final model: BETA=0.

Test	Chi-square	Df.	Pr>chsq.
Likelihood ratio	146.5312	12	<.0001
Score	140.8078	12	<.0001
Wald	142.2096	12	<.0001

The hypothesis that the model adequately fits the data can be examined by the Pearson and Deviance tests. The result reveals that the goodness of fit test is fulfilled, which is confirmed by both the deviance and chi-square with P-values of 0.4324 and 0.9696, respectively. Consequently, the observed and expected cell counts are similar and the value of each statistic is small. Moreover, in Annex 2 we check for empty cells using crosstabs with the response variable by each of the categorical predictor variables, and none of the cell are neither empty nor has too small value.

### 4.3 Interpretation of the Results

When the proportional odds model is used in the analysis of ordinal data, the coefficients of the explanatory variables in the model is interpreted as logarithm of the ratio of the odds of response variable. This means that estimates of this odds ratio, and corresponding confidence intervals, can be easily found from the fitted model. The interpretation of parameters corresponding to different variables which are found significant in the final model is described in the following section and Comparison is made with the reference category.

The log odds of severe outcome<sup>2</sup> for older age group (age between 51 and above) is decreased by 0.8508. The estimated odds ratio (OR = 0.427) indicates that the odds of serious crime (as opposed to medium or low) for older age criminals is 57% lower than young age criminals, as the odds of medium or serious crime (as opposed to low),

---

<sup>2</sup> Severe outcome stands for both serious and medium crime.

holding other variables constant. The fact that the effect of older age criminals on the odds is the same 57% for the two comparisons is one feature of the model. The odds could be as minimum as 0.323 and as maximum as 0.565 with 95% confidence.

The log odds of severe outcome for criminals who have higher education level is decreased by 0.6972. The estimated odds ratio (OR = 0.498) reveals that the odds of serious crime (as opposed to medium or low) for criminals who have higher education level is lower than those criminals who have no education by 50%. Similarly, the odds of serious or medium crime (as opposed to low) for criminals who have higher education level is lower than those criminals who have no education by 50%, holding all other variables constant. The 95% confidence interval also suggests that odds could be as minimum as 0.352 and as maximum as 0.705.

The increment to log odds of severe outcome for criminal who are government employee is 0.680. The estimated odds ratio (OR = 1.974) indicates that government employee criminals are 1.974 times more likely to commit serious crime than those criminals who are unemployed (reference group). By the same token, government employee criminals are 1.974 times more likely to attain severe outcome (to commit serious or medium crime) than unemployed criminals. The 95% confidence interval also suggests that the odds of severe crime<sup>3</sup> for government employees are 1.543 times as low and 2.525 times as high as compared to those unemployed individuals.

The log odds of a severe outcome for criminals who are self employer is increased by 0.6274. The estimated odds ratio (OR = 1.873) shows that self employer criminals are 1.873 times more likely to commit serious crime than those unemployed criminals. Likewise, self employer criminals are 1.873 times more likely to attain sever outcome (to commit serious or medium crime) than those unemployed criminals, holding all other variable constant. The 95% confidence interval also suggests that the odds of severe crime for self employer are 1.441 times as low and 2.434 times as high as compared to those unemployed criminals.

---

<sup>3</sup> Severe crime stands for both serious and medium crime.

The increment to log odds of severe outcome for criminals who are private employee is 0.606. The estimated odds ratio (OR = 1.834) reveals that private employee criminals are 1.834 times more likely to commit serious crime than those unemployed criminals. Similarly, private employee criminals are 1.834 times more likely to achieve severe outcome (to commit serious or medium crime) than unemployed criminals, keeping all other variables fixed. The odds with 95% confidence could be as minimum as 1.459 and as maximum as 2.305.

The log odds of severe outcome for criminals who are farm employee is increased by 0.467. The estimated odds ratio (OR = 1.61) indicates that farm employee criminals are 1.61 times more likely to commit serious crime than those unemployed criminals. By the same token, farm employee criminals are 1.61 times more likely to attain severe outcome (to commit serious or medium crime) than those unemployed criminals. The odds with 95% confidence could be as low as 1.297 and as high as 1.998.

The decreases in log odds of severe outcome for married criminals is 0.5934. The estimated odds ratio (OR = 0.552) implies that the odds of serious crime (as opposed to medium or low) for married criminals is lower than those unmarried criminals by 45%, as the odds of serious or medium crime (as opposed to low crime). The odds could be as minimum as 0.454 and as maximum as 0.673 with 95% confidence.

The log odds of severe outcome for urban area of crime committed is decreased by 0.2360. The estimated odds ratio (OR = 0.790) shows that the odds of serious crime (as opposed to medium or low) for urban area of crime committed is 21% lower than rural area of crime committed (reference group), as the odds of serious or medium crime (as opposed to low). The 95% confidence interval indicates that odds could be as low as 0.684 and as high as 0.912.

#### **4.4 Discussion of the Result**

The following variables: criminals age, educational level, employment status, marital status, and area of crime committed has been identified as having significant association with severe and high rate of crime so that this will help policy makers for further planning.

The study shows that there is no significant difference between the young and middle age group individuals committing crime, but there is strongly significance difference between young and older age groups. The result indicates that older age groups of individuals are less likely to commit severe crime than individuals in young age group. This implies most young age group individuals who commit crime belong to “severe crime” category and they also liable for high prevalence of crime. This is due to the fact that young age is peak age for criminal activities which provide more opportunities for committing crime. Since young individuals are less likely to be paid, own low skill to work or unemployed, they commit crime as a source of income. Uniform crime report (2009) found that the peak age (the age group with the highest age-specific arrest rate) is young age and age of offender is closely related not only to the rate at which crimes are committed but also the type of crimes committed. Thus, the finding of our study is in agreement with the Uniform crime report (2009).

Although there is a significant difference between higher education and “no education” category, there is no statistical difference between “no education” and the rest. Consequently, individuals with higher educational status have a reduced likelihood of the incidence of crime and also “severe crime”. This indicates most of uneducated individuals who commit crime belong to “severe crime” category and also they are responsible for high crime rate. A study by Lochner (2004) found that education reduce the physical returns to crime causing individuals to forego lucrative criminal opportunities, and teaches individuals to be more patient as well. The result of the study also reported that uneducated criminals committed higher crime. Thus, the finding of this study is similar to the result of our study.

Employment status also plays significant role in determining severity of crime. Among employed individuals, who commit crime, most of them belong to “severe crime” category. This is due to the nature of crime and most of severe crimes are work related crime, which includes corruption, vandalism, illegal way of import-export, and national security. Consequently, unemployed individuals have less likely to commit severe crime.

There is highly significance difference between individuals with married and single member household in terms of committing crime. That is, individuals who are married are less likely to commit “severe crime” than those unmarried. In addition, the result reveals that the overall prevalence of crime is high with those individuals who are unmarried. This may be related with different factors, such as lack of responsibility, lack of life experience, age and so on. Laub and Sampson (2003) had identified marriage as a significant force in aiding in desistance from crime and as a key turning point in facilitating desistance from crime. The finding of this study agrees with our finding.

There is highly significance difference in occurrence of crime between urban and rural. Among all crime committed in rural area, most of them belong to “severe crime” category and it also responsible for high incidence of crime. This shows people living in rural areas experience a high risk of crime than people living in urban. The possible reason for the higher frequency of crime occurred in rural area of the region due to the fact that it holds higher portion of population (80.5%), low adult literacy and low annual growth of the region (2.5%). As a result, there is high population density and have less job opportunity in rural area of the region. Additionally, this area has a smaller number of law enforcement agencies particularly police officers compared to urban. Similarly, Thakur (2003) found the causes for the growing rate of crime include unemployment, economic backwardness, over population, illiteracy and inadequate equipment of police officers.

# CHAPTER FIVE

## 5. CONCLUSION AND RECOMMENDATION

### 5.1 Conclusions

The reduced model with the logit link become the better model based on the screening criteria, the validity model assumption, the fitting statistics (Person's chi-square and deviance), and the stability of parameter estimation.

The study examined the socio-demographic and environmental determinants of crime in the region. Results of proportional odds model shows that socio-demographic and environmental variables are very important for determinant of crime outcome. The findings of the study show that different factors such as criminal's age, educational background, employment status, marital status, and area of crime committed have statistically significant effect on the outcome of crime.

The plots of outlier residuals and influential points confirmed that observations are well fitted by the model and stability of parameter estimates is fulfilled. The proportional odds model assumption was also checked numerically using score test and graphically plots of cumulative logits as parallel linear function of covariates. The results justified that the assumption was fulfilled.

### 5.2 Recommendations

The results of this study are found to be promising to be applied to address practical problems of crimes prevention in the region. It can be learn from this study that in addition to the efforts being made to reduced the frequency of crime in general, specific attention should be given to reduce the prevalence of crime by taking the following in to consideration.

- The policies and plans have to be put in place to improve young age individual education. Since young age class of the society plays a major role for development and peace of the region, they should empower through education.
- Further study can be made on the area of crime committed by considering detail and accurate information on various variables.
- All law enforcement agencies should record data in similar format and there should be presence of awareness on proper data recording system to avoid missing information (missing values) for the significant crime variables.

## Reference

\_\_\_\_\_ : “Global Report of crime”. <http://www.uncjin.org/special/globalReport.html>.

\_\_\_\_\_ : “Crime pattern”. <http://www.elon.edu/facstaff/mconklin/cis230/cases/>

Agresti, A. (2002): Categorical data analysis. John Wiley & Sons, New York, 2nd edition.

Agresti, A. A. (1996): An introduction to categorical data analysis. John Wiley & Sons, New York.

Andargachew, T. (1988): “The crime problem and its correction”. Volume I  
Addis Ababa.

Anderson, S. (1999): Crime statistics and the problem of crime in Scotland in Duff, p.  
and Hutton, N. (Eds) criminal justice in Scotland Aldershot: Aslgate.

Ary, D. Jacobs, L. C., & Razavieh, A. (1996): Introduction to research in education. Fort Worth, TX: Harcourt Brace College Publishers.

Bennell, C., & Canter, D.V. (2002): “Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis”. Science and Justice, **42**, 153-164.

Brown, D. (2003): The Regional Crime Analysis Program (RECAP). A  
Framework for mining data to catch criminals.  
<http://vijis.sys.virginia.edu/publication/RECAP.pdf>

Chiricos, T. (1987): “Rates of Crime and Unemployment: An Analysis of Aggregate Research Evidence”. Social Problems, **34**, 187-212.

Cochran, W. G. (1977): Sampling techniques. John Wiley & Sons, New York, 3rd

Demographic, E. (2005): "Health survey (DHS)". Central Statistics Authority and ORC  
Marc, Addis Ababa, Ethiopia and Calverton, Maryland. USA.

Dinkes, R. (2004): Neighbourhoods Matter: Fixed and Random effects in police satisfaction surveys in the UK. Masters thesis, London school of economic and political science, methodology institute and department of statistics, UK.

FDREPCC, (2008): "Summary and statistical report of the 2007 population size by age and sex". Central Statistical Agency, Addis Ababa, Ethiopia.

Federal Bureau of investigation Uniform Crime Report (2009): Crime in United States 2008.

Federal police commission, (2011): national crime statistics, Addis Ababa, Ethiopia.

Fink, A. (1995): The survey handbook. Thousand Oaks, CA: Sage Publications.

FMOH (2005): Health Sector Strategic Plan (HSD-III) 2005/6-2009/10, Federal ministry of Health, Addis Ababa.

Freeman, R. (1996): "Why do so many young American men commit crimes and what might we do about it?". Journal of Economic Perspectives **10**, 25-42.

Freeman, R. B. (1983): "Crime and Unemployment". Crime and Public Policy, ICS Press, 89-106, San Fransisco.

Gotfredson, M. and Hirschi, T. (1996): A general theory of crime. Stanford university press, Stanfored, California.

- Harrell, Jr., F. E. (2001): Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis. Springer-Verlag, New York.
- Hosmer, D. W. and Lemeshow, S. (2000): Applied logistic regression, John Wiley & Sons, New York, 2nd edition.
- International City Managers association, (1961): Municipal Police Administration. USA, Cushing Malloy, Inc and Ann Arbor, Michigan.
- Davis, B. J. (2006): "Distribution of Property Crime and Police Arrest Rates across Los Angeles Neighborhoods". Western Criminology Review, **7**(3), 7-26.
- Krejcie, R. V., & Morgan, D. W. (1970): "Determining sample size for research activities". Educational and Psychological Measurement, **30**, 607-610.
- Laub, John H., and Robert J. S. (2003): Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70. Cambridge: Harvard University Press.
- Lochner, L. (2004): "Education, work, and crime: a human capital approach". International Economic Review, **45**, 811-43.
- Machin, S., and Meghir, C. (2004): "Crime and economic incentives". The Journal of Human Resources **39**, 958-79.
- Megaputer Intelligence, Inc. (2002): Crime Pattern Analysis. Megaputer Case Study. [http://www.elon.edu/facstaff/mconklin/cis230/cases/crime\\_pattern\\_case.pdf](http://www.elon.edu/facstaff/mconklin/cis230/cases/crime_pattern_case.pdf)
- Mesfin, G/M. (1999): "A computer based criminal records management system case study of the Addis Ababa police". Unpublished Master's Thesis, Addis Ababa University SISA, Addis Ababa.
- Osgood, D. Wayne. 2000: "Poisson-Based Regression Analysis of Aggregate Crime Rates". Journal of Quantitative Criminology, **16**(1), 21-43.

- Salkind, N. J. (1997): Exploring research. Upper Saddle River, NJ: Prentice Hall, 3rd edition.
- Sampson, R., & W.J. Wilson. (1995): Toward a theory of race, crime & urban inequality. In J. Hagan and R.D. Peterson (eds.) Crime and Inequality. Stanford, CA: Stanford University press.
- SAS Institute, Inc. (2004): SAS/STAT 9.1 user's guide. SAS Institute Inc., Cary, NC.
- Thakur, C. (2003): "Crime Control". <http://ncthakur.itgo.com/chand3c.htm>.
- Tseloni, A. (1999): "Personal criminal victimization in the U.S.: Fixed and random effects of individual and Household characteristics". British criminology conference, Liverpool, 12-16.
- U.S congress, House committee on the judiciary, subcommittee on crime, unemployment and crime, 95<sup>th</sup> cong., 1<sup>st</sup> sess., September 27, 1978.
- Wilson, J.Q. and Kelling, G. (1982): "Broken Windows: The Police and Neighborhood Safety". Atlantic Monthly, 29-38.
- Wilson, O.W. (1963): Police Administration. USA, McGraw Hill Company.
- Wilson, p., Quisenberry, N. and Jones, S. (2003): "The built environment and community risk interpretation". Journal of research in crime and delinquency, **40**(3), 322.

## Annex 1. SAS out Put

### A.1.1 Logistic regression Full Model

Probabilities modeled are cumulated over the lower Ordered Values.

#### Score Test for the Proportional Odds Assumption

Chi -Square	DF	Pr > Chi Sq
14. 6924	13	0. 3269

#### Deviance and Pearson Goodness-of-Fit Statistics

Cri teri on	Val ue	DF	Val ue/DF	Pr > Chi Sq
Devi ance	682. 2498	653	1. 0448	0. 2073
Pearson	607. 2548	653	0. 9299	0. 8994

#### Model Fit Statistics

Cri teri on	I ntercept Only	I ntercept and Covari ates
AIC	5972. 303	5849. 587
SC	5984. 144	5938. 394
-2 Log L	5968. 303	5819. 587

#### Testing Global Null Hypothesis: BETA=0

Test	Chi -Square	DF	Pr > Chi Sq
Li kel i hood Rati o	148. 7161	13	<. 0001
Score	142. 9334	13	<. 0001
Wal d	144. 1605	13	<. 0001

#### Type 3 Analysis of Effects

Effect	DF	Chi -Square	Pr > Chi Sq
Sex	1	2. 1681	0. 1409
Age	2	37. 9355	<. 0001
Edu. cat.	4	15. 9142	0. 0031
Emp.	4	42. 5548	<. 0001
Mar	1	35. 2156	<. 0001
Area	1	9. 8549	0. 0017

Table (4) Analysis of Maximum Likelihood and Odds Ratio Estimates

Parameter	estimate	Standard error	Wald Chi-square	Df.	p-value	exp ( $\beta$ )	95% Wald Confidence limits
Intercept=0	-0.2704	0.0856	9.9747	1	.0016	-	-
Intercept=1	1.1830	0.0887	177.8932	1	<.0001	-	-
Female	0.2023	0.1374	2.1681	1	0.1409	1.224	0.935-1.603
Age(51 and above)	-0.8602	0.1434	35.9778	1	<.0001	0.423	0.319-0.560
Higher education	-0.6983	0.1775	15.4805	1	<.0001	0.497	0.351-0.704
Government emp.	0.6851	0.1257	29.7179	1	<.0001	1.984	1.551-2.538
self employ	0.6255	0.1339	21.8338	1	<.0001	1.869	1.438-2.430
Privet employee	0.6084	0.1167	27.1736	1	<.0001	1.838	1.462-2.310
Farm employee	0.4819	0.1103	19.0914	1	<.0001	1.619	1.304-2.010
married marital status	-0.5969	0.1006	35.2156	1	<.0001	0.551	0.452-0.671
Rural area	-0.2314	0.0737	9.8549	1	0.0017	0.793	0.687-0.917

## A.1.2 Logistic Regression Reduced model

### Model Information

Data Set	WORK. CRIME
Response Variable	crime severity
Number of Response Levels	3
Model	cumulative Logit
Optimization Technique	Fisher's scoring

### Response Profile

Ordered Value	crime severity	Total Frequency
1	0	1118
2	1	901
3	2	734

### Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > Chi Sq
14.5882	12	0.2647

### Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > Chi Sq
Deviance	472.5561	468	1.0097	0.4324
Pearson	412.3540	468	0.8811	0.9696

### Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	5972.303	5849.772
SC	5984.144	5932.658
-2 Log L	5968.303	5821.772

R-Square 0.0518      Max-rescaled R-Square 0.0585

### Type 3 Analysis of Effects

Effect	DF	Chi-Square	Pr > Chi Sq
Age	2	37.2345	<.0001
Edu. cat.	4	16.0101	0.0030
Emp.	4	42.2417	<.0001
Mar.	1	34.8514	<.0001
Area	1	10.2605	0.0014

Table (5) Analysis of Maximum Likelihood and Odds Ratio Estimates

Parameter	Estimate	Standard error	Wald Chi-square	DF.	p-value	$\exp(\beta)$	95% Wald CL	
Intercept=0	-0.2487	0.0844	8.6924	1	0.0032	-	-	
Intercept=1	1.2030	0.0876	188.6896	1	<.0001	-	-	
Age(51 and above)	-0.8508	0.1432	35.3043	1	<.0001	0.427	0.323	0.565
Higher education	-0.6972	0.1774	15.4358	1	<.0001	0.498	0.352	0.705
Government	0.6801	0.1256	29.3102	1	<.0001	1.974	1.543	2.525
self employ	0.6274	0.1338	21.9717	1	<.0001	1.873	1.441	2.434
Privet employee	0.6064	0.1167	27.0002	1	<.0001	1.834	1.459	2.305
Farm employee	0.4761	0.1102	18.6708	1	<.0001	1.610	1.297	1.998
Married marital status	-0.5934	0.1005	34.8514	1	<.0001	0.552	0.454	0.673
urban	-0.2360	0.0737	10.2605	1	0.0014	0.790	0.684	0.912

### A.1.3 Interaction model fit for full and reduced model

#### **Full model with interaction result**

Score Test for the Proportional Odds Assumption

Chi -Square	DF	Pr > Chi Sq
99.3935	78	0.0516

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	5972.303	5904.478
SC	5984.144	6378.114
-2 Log L	5968.303	5744.478

#### **Reduced model with interaction result**

Score Test for the Proportional Odds Assumption

Chi -Square	DF	Pr > Chi Sq
89.0712	65	0.0254

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	5972.303	5894.460
SC	5984.144	6291.129
-2 Log L	5968.303	5760.460

#### **Type 3 Analysis of Effects**

Effect	DF	Chi -Square	Pr > Chi Sq
Age	2	12.3302	0.0021
Edu. cat.	4	4.6277	0.3277
Emp.	4	20.2478	0.0004
Mar.	1	2.2999	0.1294
Area	1	1.0613	0.3029
age*educat.	8	10.9248	0.2060
age*emp.	8	12.8495	0.1171
age*mar.	2	1.1822	0.5537
age*area	2	6.2546	0.0538
educat*emp.	16	14.1349	0.5887
educat*mar.	4	7.9376	0.0939
educat*area	4	0.3733	0.9846
emp*mar.	4	1.4275	0.8394

## Annex 2: Assessing of Empty Cell

Table 6: Criminal sex \* crime severity Crosstab

Sex category of criminals		Crime severity			Total
		serious	medium	low	
Male	count	1027	837	683	2574
Female	count	91	64	51	206
Total	count	1118	901	734	2753

Table 7: Criminal age \* crime severity Crosstab

Age category		Crime severity			Total
		serious	medium	low	
12-30	Count	740	528	389	1657
31-50	Count	323	291	236	850
51 & above	Count	55	82	109	246
Total	Count	1118	901	734	2753

Table8: Criminal Education level \* crime severity Crosstab

		Crime severity			Total	
		serious	Medium	low		
Education	No education	Count	544	418	321	1283
	Primary edu.	Count	225	186	148	559
	Junior edu.	Count	164	139	101	404
	Secondary edu.	Count	145	130	107	382
	Higher edu.	Count	40	28	57	125
Total	Count		1118	901	734	2753

Table 9: Criminal employment status \* crime severity Crosstab

			Crime severity			Total
			serious	medium	low	
employment	Gov't	Count	172	137	107	416
	Self emp.	Count	128	110	72	310
	Priv.emp.	Count	215	171	131	517
	Farm.emp	Count	284	206	192	682
	Unemp.	Count	319	277	232	828
Total		Count	1118	901	734	2753

Table 10: Criminal marital status \* Crime severity Crosstab

			Crime severity			Total
			serious	medium	Low	
Marital	single	Count	691	482	338	1511
	married	Count	427	419	396	1242
Total		Count	1118	901	734	2753

Table 11: Area of crime committed \* crime severity Crosstab

			Crime severity			Total
			Serious	medium	Low	
Area	Rural	Count	711	540	416	1667
	Urban	Count	407	361	318	1086
Total		Count	1118	901	734	2753

## Annex 3: Descriptive Statistics

Table 13: Criminals education level \* criminal employment status Crosstab

Education category		Employment category					Total
		Government employee	Self employ	Privet employee	Farm employee	Un employed	
No education	Count	165	136	224	406	352	1283
	%	12.9%	10.6%	17.5%	31.6%	27.4%	100.0%
Primary education	Count	81	48	100	137	193	559
	%	14.5%	8.6%	17.9%	24.5%	34.5%	100.0%
Junior secondary education	Count	59	61	99	60	125	404
	%	14.6%	15.1%	24.5%	14.9%	30.9%	100.0%
Secondary education	Count	68	40	76	60	138	382
	%	17.8%	10.5%	19.9%	15.7%	36.1%	100.0%
Higher education	Count	43	25	18	19	20	125
	%	34.4%	20.0%	14.4%	15.2%	16.0%	100.0%
Total	Count	416	310	517	682	828	2753
	%	15.1%	11.3%	18.8%	24.8%	30.1%	100.0%

Table 14: percentage of variables

variable	Levels	Number of crime	percentage
Criminal Age	12-30	1657	60.2%
	31-50	850	30.9%
	51 and above	246	8.9 %
	total	2753	100%
Criminal Edu.	No education	1283	46.6%
	Primary educa.	559	20.3%
	Junior Secondary	404	14.7%
	Secondary educa.	382	13.9%
	Higher education	125	4.5%
	Total	2753	100%
Criminal Emp. status	Gov't. employee	416	15.1%
	Self employ	310	11.3%
	Privet employee	517	18.8%
	farm employee	682	24.8%
	Un employee	828	30.1%
	Total	2753	100%
Criminal Marital status	Single	54.9	54.9%
	Married	45.1	45.1%
	Total	2753	100%
Area of Crime Committed	Rural	1667	60.6%
	Urban	1086	39.4%
	Total	2753	100%

Table 15: Total number of crime in Ethiopia reported to police within three consecutive years.

Year	Number of criminals			Crime committed	Remark
	Male	Female	Total		
2007/08	280,376	36,351	317,727	195,608	Somalia region is not included
2008/09	283,013	37,172	320,185	198,250	Harar region is not included
2009/10	280,601	36,270	316,871	211,202	SNNPR is not included

Table 16: Total number of crime in Tigray region reported to police within three consecutive years.

Year	Number of criminals			Crime committed
	Male	Female	Total	
2007/08	38,514	5,685	44,199	23,232
2008/09	21,885	3,408	25,293	16,248
2009/10	35,710	5,034	40,744	27,940

Annex 4. Figure for cumulative logits represented as parallel and linear fuhnction of variables

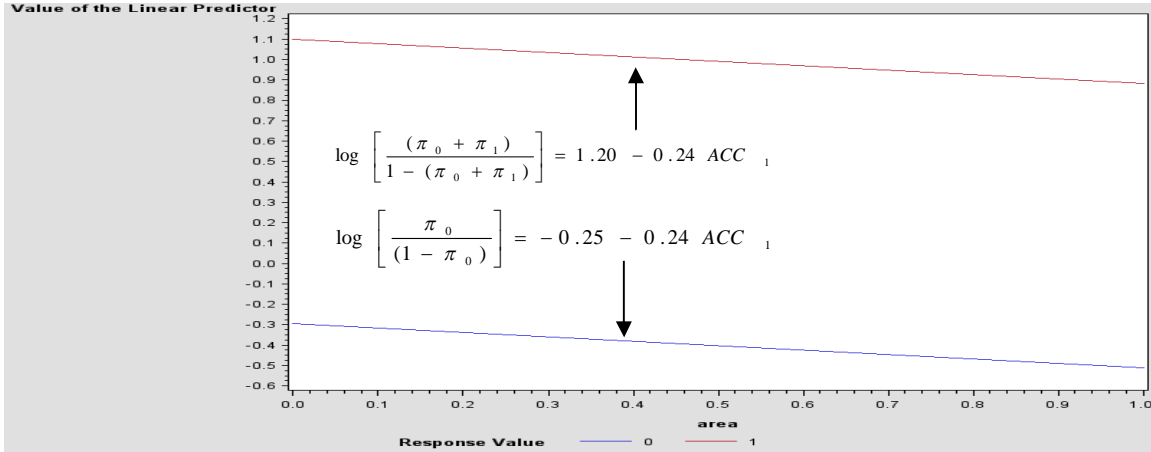


Figure 8: polts of cumulative logits of crime severity represented as Parallel linear function of variable Area.

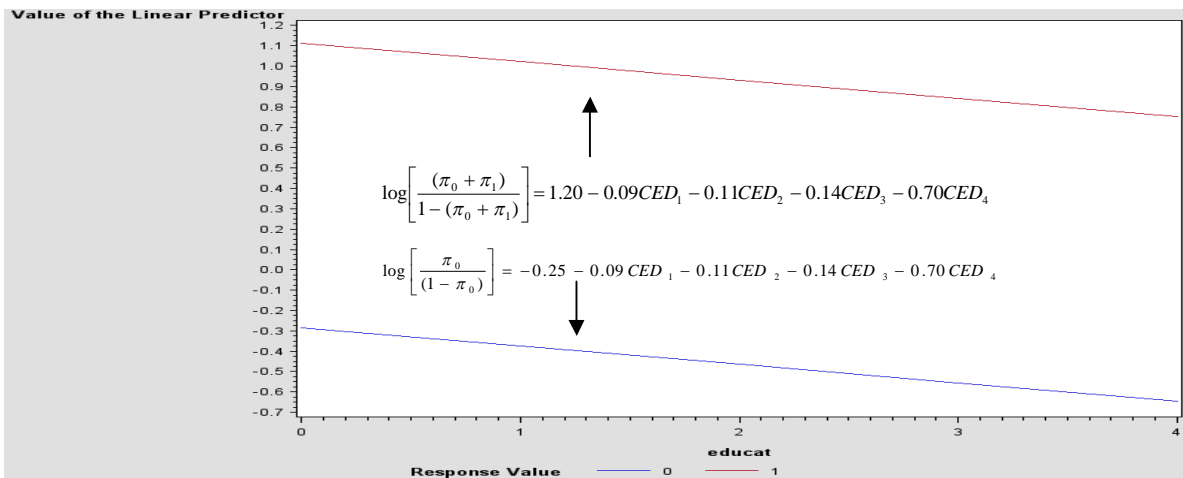


Figure 9: polts of cumulative logits of crime severity represented as Parallel linear function of variable Education.

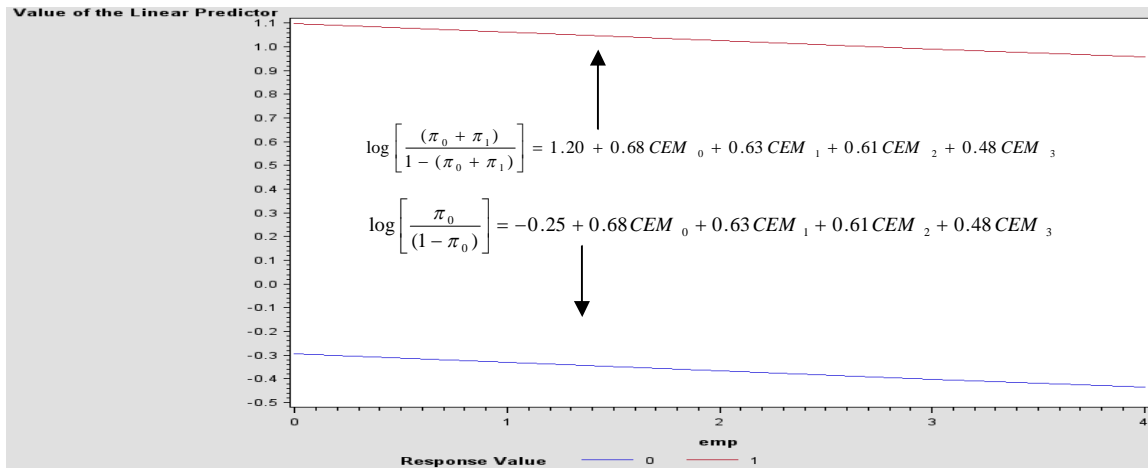


Figure 10: polts of cumulative logits of crime severity represented as Parallel linear function of variable Employment.

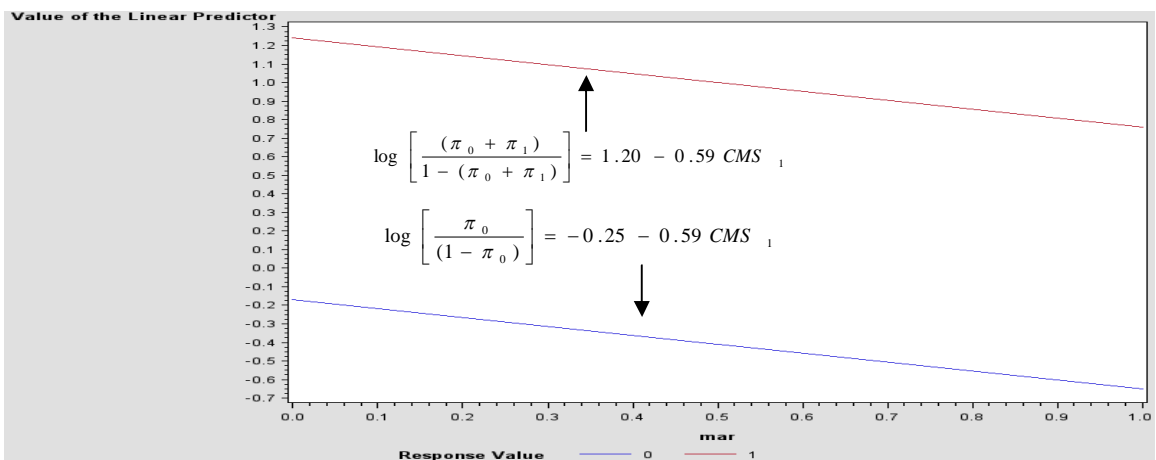


Figure 11: polts of cumulative logits of crime severity represented as Parallel linear function of variable Marital status.

## Annex 5: About Type 3 Analysis of SAS

A Type 3 analysis is similar to the Type III sums of squares used in PROC GLM, except that likelihood ratios are used instead of sums of squares. First, a Type III estimable function is defined for an effect of interest in exactly the same way as in PROC GLM. Then, maximum likelihood estimation is performed under the constraint that the Type III function of the parameters is equal to 0, using constrained optimization. Let the resulting constrained parameter estimates be  $\tilde{\beta}$  and the log likelihood be  $l(\tilde{\beta})$ . Then the likelihood ratio statistic is

$$S = 2 \left\{ l(\hat{\beta}) - l(\tilde{\beta}) \right\}$$

where  $\hat{\beta}$  is the unconstrained estimate, has an asymptotic chi-square distribution under the hypothesis that the Type III contrast is equal to 0, with degrees of freedom equal to the number of parameters associated with the effect.

In PROC LOGISTIC a Type 3 analysis produces a table that contains the likelihood ratio statistics, degrees of freedom, and  $p$ -values based on the limiting chi-square distributions for each effect in the model.

## Annex 6: SAS Codes

```
Proc logistic data = crime;  
Class sex(ref='0')age(ref='0') educat(ref='0') emp(ref='4')  
mar(ref='0') area(ref='0') /param= ref order = internal;  
model crimeseverity = sex age educat emp mar area /link =  
clogit scale = none aggregate;  
format sex age educat emp mar area ;  
run;
```

---

```
Proc logistic data = crime;  
Class age(ref='0')educat(ref='0') emp(ref='4') mar(ref='0')  
area(ref='0') /param= ref order = internal;  
model crimeseverity = sex age educat emp mar area /link =  
clogit scale = none aggregate;  
format age educat emp mar area ;  
run;
```

---

```
ods graphics on;  
ods html;  
Proc logistic data = crime;  
model crimeseverity = age educat emp mar area/link =  
clogit;  
output out = pred p=p xbeta=linp;  
run;  
proc sort data = pred;  
by age;  
run;  
options reset = all ;  
symbol i = join w=.4 ;  
proc gplot data = pred;  
plot p*age=_level_;  
plot linp*age=_level_;  
run;  
quit;  
ods html close;  
ods graphics off;
```

---

```
Proc logistic data = crime ;  
class sex(ref='0')age(ref='0') educat(ref='0') emp(ref='4')  
mar(ref='0') area(ref='0') /param= ref order = internal;  
model crimeseverity = sex age educat emp mar area  
age*educat age*emp age*mar age*area /link = clogit scale =  
none aggregate;  
format sex age educat emp mar area ;  
run;
```

---

## DECLARATION

I, the undersigned, declare that the thesis is my original work, has not been presented for degrees in any other university and all sources of material used for the thesis have been duly acknowledged.

Name: \_\_\_\_\_

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

This thesis has been submitted for examination with my approval as a University Advisor.

.....

Dr.Grima Taye