MICRO-LEVEL ANALYSIS OF UNREGISTERED EMPLOYMENT IN TURKEY WITH GROUP COMPARISONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MEHMET İNER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2019

Approval of the thesis:

MICRO-LEVEL ANALYSIS OF UNREGISTERED EMPLOYMENT IN TURKEY WITH GROUP COMPARISONS

submitted by **MEHMET İNER** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşen D. Akkaya
Head of Department, **Statistics** _____

Prof. Dr. Ayşen D. Akkaya
Supervisor, **Statistics, METU** _____

Assoc. Prof. Dr. Özlem Türker Bayrak
Co-Supervisor, **Inter-Curricular Courses Dept., Çankaya Uni.** _____

Examining Committee Members:

Assoc. Prof. Dr. Ebru Yüksel Haliloğlu
Management, TOBB University of Economics and Technology _____

Prof. Dr. Ayşen D. Akkaya
Statistics, METU _____

Assoc. Prof. Dr. Ceylan Yozgatlıgil
Statistics, METU _____

Date: 05.09.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Mehmet İner

Signature:

ABSTRACT

MICRO-LEVEL ANALYSIS OF UNREGISTERED EMPLOYMENT IN TURKEY WITH GROUP COMPARISONS

İner, Mehmet
Master of Science, Statistics
Supervisor: Prof. Dr. Ayşen D. Akkaya
Co-Supervisor: Assoc. Prof. Dr. Özlem Türker Bayrak

September 2019, 111 pages

Group comparison of logistic regression models in a similar way with OLS is manipulating depending on the unobserved heterogeneity in logistic regression. In this sense, this study focuses on the group comparison problem in logistic regression. In order to get to the root of the comparison problem in logistic regression, the theoretical background of the logistic regression is explained with the latent propensity interpretation in which the extent of the dependent variable's closeness to success is taken into consideration. In this respect, the discussions on the diagnosis and the remediation of the problem in the literature are revealed and analyzed. The application of group comparison in logistic regression is made by means of unregistered employment data in Turkey. In this respect, comparisons among genders, regions and years are made in terms of unregistered employment in this thesis. For this aim, Long's (2009) and Long & Mustillo's (2018) methods to conduct comparisons among groups by means of predicted probabilities and marginal effects are utilized since the test of difference in predicted probabilities based on the models and the marginal effects are not scaled by unobserved heterogeneity. Various important socio-economical results and implications including gender differences and regional differences are reached through the comparisons. Moreover, the differences in marginal effects in 10 years

period are analyzed and the changes over time are associated with the measures taken in the field.

Keywords: Logistic Regression, Latent Propensity, Unobserved Heterogeneity, Unregistered Employment, Micro Level Analysis

ÖZ

TÜRKİYE'DE KAYIT DIŞI İSTİHDAMIN GRUPLAR ARASI KIYASLAMALAR YOLUYLA MİKRO DÜZEYLİ ANALİZİ

İner, Mehmet
Yüksek Lisans, İstatistik
Tez Danışmanı: Prof. Dr. Ayşen D. Akkaya
Ortak Tez Danışmanı: Doç. Dr. Özlem Türker Bayrak

Eylül 2019, 111 sayfa

Lojistik regresyon modellerinde OLS ile benzer şekilde gruplar arası karşılaştırma yapılması, lojistik regresyondaki gözlemlenemeyen heterojenliğe bağlı olarak yanıltıcı sonuçlar verir. Bu nedenle bu çalışma lojistik regresyonda gruplar arası karşılaştırma problemine odaklanmaktadır. Lojistik regresyonda karşılaştırma probleminin kökenine ulaşmak için, lojistik regresyonun teorik arka planı, bağımlı değişkenin gerçekleşmeye yakınlık derecesinin dikkate alındığı gizli eğilim (latent propensity) yorumlaması vasıtasıyla açıklanmaktadır. Bu bağlamda, literatürde sorunun teşhisi ve giderilmesine ilişkin tartışmalar ortaya konulmuş ve analiz edilmiştir. Bu çalışmada lojistik regresyonda grup karşılaştırması uygulaması, Türkiye'deki kayıt dışı istihdam verileri yoluyla yapılmaktadır. Bu bağlamda, bu tezde kayıtdışı istihdam açısından cinsiyetler, bölgeler ve yıllar arasında karşılaştırmalar yapılmıştır. Bu amaçla, modeller aracılığıyla tahmin edilen olasılıklar ve marjinal etkiler ile gruplar arasında karşılaştırma yapmayı öneren Long (2009) ve Long & Mustillo (2018) yöntemleri uygulanmıştır. Bu yöntem, modeller aracılığıyla tahmin edilen olasılıkların ve marjinal etkilerin gözlem dışı heterojenlikten etkilenmemesi sebebiyle tercih edilmiştir. Karşılaştırmalar vasıtasıyla cinsiyet farklılıkları ve bölgesel farklılıklar dahil olmak üzere çeşitli önemli sosyo-ekonomik sonuçlar elde

edilmektedir. Ayrıca, 10 yıllık dönemdeki marjinal etki farklılıkları analiz edilerek ve zaman içindeki değişimler bu alanda alınan önlemlerle ilişkilendirilmektedir.

Anahtar Kelimeler: Lojistik Regresyon,Gizli Eğilim, Gözlem Dışı Heterojenlik, Kayıt Dışı İstihdam, Mikro Düzeyli Analiz

To Ahmet Aras

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Prof. Dr. Ayşen D. Akkaya and co-supervisor Assoc. Prof. Dr. Özlem Türker Bayrak for their guidance, comments and suggestions during the course of the project.

I would like to present my grateful thanks to my examing committee members, Assoc. Prof. Dr. Ebru Yüksel Haliloğlu and Assoc. Prof. Dr. Ceylan Yozgatlıgil for reviewing my study.

I am grateful to my wife. The completion of this undertaking could not have been possible without her support. Finally, my sincere thanks to my family for their support and motivation during all these years.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGEMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xv

LIST OF FIGURES xvi

LIST OF ABBREVIATIONS xviii

CHAPTERS

1. INTRODUCTION 1

1.1. Contribution to the Literature 3

2. LOGISTIC REGRESSION MODELS 5

2.1. Linear Regression Models 5

2.2. Generalized Linear Models 6

2.3. Logistic Regression Models 7

2.3.1. Interpretation of Logistic Regression Models 10

2.3.2. Model Fitting in Logistic Regression via Maximum Likelihood Estimation 12

2.3.3. Significance Testing in Logistic Regression 14

2.3.4. Goodness of Fit in Logistic Regression 15

2.3.4.1. Pearson Chi-squared Test 16

2.3.4.2. Deviance Test 17

2.3.4.3. The Hosmer-Lemeshow Test 17

2.3.4.4. Classification Tables 18
2.3.4.5. Area Under ROC Curve 19
2.3.4.6. Pseudo **R2** Test 20
3. GROUP COMPARISON PROBLEM IN LOGISTIC REGRESSION 21
3.1. Latent Propensity Perspective in Logistic Regression 21
3.2. Group Comparison Problem in Logistic Regression 23
4. UNREGISTERED EMPLOYMENT 29
4.1. Informal Sector 29
4.2. Unregistered Employment 30
4.2.1. Reasons of Unregistered Employment 32
4.2.1.1. Economic Reasons 32
4.2.1.2. Social Reasons 33
4.2.1.3. Reasons Arising from Labour Market Status 33
4.2.2. Impacts of the Unregistered Employment 34
4.2.2.1. Financial Impacts 34
4.2.2.2. Economic Impacts 34
4.2.2.3. Impacts on Employees 35
4.2.3. Fighting the Unregistered Employment 35
4.2.4. The Unregistered Employment in Turkey 36
4.2.5. Characteristics of Unregistered Employment in Turkey 38
4.2.5.1. Gender 38
4.2.5.2. Age 39
4.2.5.3. Education Status 41
4.2.5.4. Marital Status 42

4.2.5.5. Sector 42
4.2.5.6. Number of Employees Employed in the Workplace 43
4.2.5.7. Employment Status 44
4.2.5.8. Type of Employment.....44
4.2.5.9. Region 45
4.2.6. Fighting the Unregistered Employment in Turkey 46
5. ANALYSIS OF THE UNREGISTERED EMPLOYMENT IN TURKEY 49
5.1. Aims of Modelling 49
5.2. Review of the Previous Similar Studies in Literature 50
5.3. Household Labour Force Survey.....53
5.4. Information about the Variables.....54
5.5. Econometric Analysis.....58
5.5.1. Analysis of the Overall Models for 2007 and 2017.....59
5.5.1.1. Revealing the Differences among 2007 and 2017 64
5.5.2. Analysis of Separate Models According to Gender.....66
5.5.2.1. Revealing the Differences among Genders.....70
5.5.3. Analysis of Separate Models According to Region.....74
5.5.3.1. Revealing the Differences among Regional Groups.....78
6. CONCLUSION.....81
REFERENCES.....85
APPENDICES
A. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for 2017 91

B. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for 2007 93

C. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Males 95

D. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Females 97

E. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Eastern Regions 99

F. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Western Regions 101

G. R Codes Used in the Analysis 103

LIST OF TABLES

TABLES

Table 4.1. Unregistered Employment Numbers and Rates in Turkey (1988 - 2018) 37
Table 5.1. Categories of the Region Independent Variable and the Provinces under each category.....56
Table 5.2. AME's of the Micro-Factors of Unregistered Employment for 2007 and 2017.....62
Table 5.3. AME's of the Micro-Factors of Unregistered Employment for Males and Females 69
Table 5.4. AME's of the Micro-Factors of Unregistered Employment for Eastern and Western Regions 77
Table 5.5. Probabilities of Unregistered Employment for Different Regions by Different Levels of Education and Different Sectors.....79

LIST OF FIGURES

FIGURES

Figure 2.1. Logistic Regression Function 9
Figure 2.2. Area Under ROC Curve 19
Figure 4.1. The Trend of Unregistered Employment in Turkey (Turkstat, HLFS, 1988 - 2018 37
Figure 4.2. The Difference of Unregistered Employment Rates by Genders (ILO, 2018) 39
Figure 4.3. Global Registered Employment Rates by Different Age Groups (ILO, 2018) 39
Figure 4.4. Histogram of Unregistered Employment Rates by Different Age Groups (Turkstat, HLFS, 2017) 40
Figure 4.5 Global Registered Employment Rates by Different Levels of Education (ILO, 2018) 41
Figure 4.6. Unregistered Employment Rates by Different Levels of Education in Turkey as of 2017 (Turkstat, HLFS, 2017) 42
Figure 4.7. Unregistered Employment Rate by Numbers of Employees in the Work Place (Turkstat, HLFS, 2017) 43
Figure 4.8. Unregistered Employment Rates by Different Employment Status as of 2017 (Turkstat, HLFS, 2017) 44
Figure 4.9. Unemployment Rates by NUTS-II Regions (Turkstat, HLFS, 2017) 45
Figure 5.1. Accuracy Rate – Cut-off Point Graphs for the Overall Models of 2007(a) and 2017(b) 60
Figure 5.2. ROC Curves for the Overall Models of 2007(a) and 2017(b) 60
Figure 5.3. Accuracy Rate – Cut-off Point Graph for the Models of Male(a) and Female(b) Groups 66
Figure 5.4. ROC Curve for the Models of Male(a) and Female(b) Groups 67

Figure 5.5. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Non-Educated Individuals.....71
Figure 5.6. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Primary School Graduates.....72
Figure 5.7. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Secondary School Graduates.....72
Figure 5.8. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for High School Graduates.....73
Figure 5.9. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Vocational High School Graduates 73
Figure 5.10. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Graduates of Higher Education and More.....74
Figure 5.11. Accuracy Rate – Cut-off Point Graph for the Models of East(a) and West(b).....75
Figure 5.12. ROC Curves for the Models of East and West.....76

LIST OF ABBREVIATIONS

ABBREVIATIONS

AME	Average Marginal Effect
CDF	Cumulative Distribution Function
EU	European Union
FN	False Negative
FP	False Positive
GLM	Generalized Linear Models
HLFS	Household Labour Force Survey
ILO	International Labour Organization
LR	Likelihood Ratio
MER	Marginal Effect at Representative Values
MLE	Maximum Likelihood Estimation
NACE	Statistical Classification of Economic Activities in the European Community
NUTS	Nomenclature of Territorial Units for Statistics
OLS	Ordinary Least Square Method
PDF	Probability Density Function
PMF	Probability Mass Function
ROC	Receiver Under Receiver Operating Characteristic
SE	Standard Error

SME	Small Medium Sized Entreprises
SSI	Social Security Institution
TN	True Negative
TP	True Positive

CHAPTER 1

INTRODUCTION

Logistic regression is a widely accepted statistical tool when the response variable is dichotomous. Especially in economics, there are many cases in which the researchers focus on observing the determinants of variables with binary characteristics. However, the binary nature of the response variable in logistic regression may be considered as strict in some researches as it does not let the dependent variable to get a value between 0 and 1. In other words, the researcher may seek for the dependent variable's extent of closeness to 0 or 1.

In this context, the latent propensity perspective in logistic regression assumes an unobservable, hypothetical y^* value which is in a linear relationship with the independent variables. The change in the latent propensity leads to a change in the observed response variable after a certain threshold.

The comparison of the independent variables' effects on the response variable for different groups is problematic in logistic regression. Allison (1999) started the discussion on this issue by means of the latent propensity perspective and indicated that directly comparing the coefficients as in OLS is manipulating depending on the unobserved heterogeneity arising from the non-observed or omitted covariates.

Subsequent to the detection of group comparison problem in logistic regression analysis, several methods have been put forward to overcome, some of which are contradictory with each other. In this respect, this thesis study focuses on Long's (2009) and Long & Mustillo's (2018) methods to conduct comparisons among groups by means of predicted probabilities and marginal effects to probabilities since the test of difference in predicted probabilities based on the models and the marginal effects are not scaled by unobserved heterogeneity. Furthermore, group comparison based on

predicted probabilities manages to indicate the group differences for different levels of the same independent variable and enables the researcher to construct confidence interval for the difference of predicted probabilities.

In this thesis, the application of group comparison in logistic regression is made by means of unregistered employment data in Turkey. In this respect, the unregistered employment is approached as working without any social security relating to the main job in line with Turkstat's Household Labour Force Survey (HLFS), even though there are various different definitions in other sources.

Unregistered employment is a salient phenomenon in Turkey. Depending on the economic reasons including unemployment, inflation and sectoral distribution, the social reasons including income inequality, poverty, immigration and population increase and the reasons arising from the labour market status, unregistered employment is very common in Turkey. According to Turkstat's HLFS, unregistered employment rate is 33,4% and non-agricultural unregistered employment rate is 22.3% in 2018. Considering the unregistered employment as an important problematic area in Turkish labour market with various financial, economic and social impacts, it is assessed as crucial to conduct an individual based micro analysis of the phenomenon in order to provide information for the policies developed and measures taken.

Unregistered employment is an appropriate field of study in order to conduct group comparison methods in logistic regression. In this respect, comparisons among years, genders, and regions are made in terms of unregistered employment in this thesis. Accordingly, various important socio-economical results and implications including gender differences and regional differences are reached. Moreover, the differences in marginal effects in 10-year period are analyzed and some of the changes over time are associated with the measures taken in the field.

The structure of the thesis is as follows. Chapter 2 starts with the theoretical description of the logistic regression with its place and its function in the field of regression models. This chapter briefly includes interpretation of logistic regression

as well as model fitting, significance testing and goodness of fit in logistic regression. Chapter 3 underlines the latent propensity perspective in logistic regression prior to putting forward the group comparison problem in logistic regression. Afterwards, it sheds light on the discussions in the literature and overcoming methods proposed regarding the problem. Subsequently, Chapter 4 provides thematic information about the unregistered employment including its reasons, impacts, characteristics and status in Turkey. Then, Chapter 5 includes the application of logistic regression to unregistered employment data in Turkey and comparisons of groups and detailed econometric analysis of the models and comparisons. Finally, Chapter 6 sums up the study and puts forth the remarkable results.

1.1. Contribution to the Literature

Unregistered employment is a frequent research topic both in Turkey and in the world. In this context, most of the researches associate unregistered employment with the macro-factors. In comparison with the studies focusing on the macro-factors, there are limited number of micro-level studies in this field. However, it should be noted that, while the general rate of unregistered employment is strongly related with macro-factors, the risk of a specific person to take place in unregistered employment is also determined by micro-factors based on the personal specifications.

In addition to the fact that there are few numbers of studies regarding the micro-level analysis of unregistered employment in Turkey, the contents of these micro-level studies are limited in terms of commonly accepted features. To be more precise, it is theoretically well known that there are gender differences, regional disparities, and differences by time in this field. However, these differences are not usually attributed to improved statistical methods. In this context, this thesis provides an opportunity to take a closer look at the gender differences, in terms of factors of unregistered employment based on personal specifications. Moreover, it makes comparisons

between regions and reveals several remarkable differences. Also, it observes the changes in micro-factors by time.

To summarize, it can be argued that this study, with its comparison perspective after a comprehensive review of the statistical discussions in the field, includes a further level of analysis compared to the other similar micro-level studies within the field of unregistered employment.

CHAPTER 2

LOGISTIC REGRESSION MODELS

This chapter overviews the concept of Logistic Regression Models in connection with the Generalized Linear Models (GLM) with an holistic approach. Afterwards, it focuses on the interpretation of the Logistic Regression Model as well as the procedures of model fitting, significance testing and goodness of fit.

2.1. Linear Regression Models

Regression analysis is a widely used statistical method within the scope of data analysis. It investigates the relationship between variables in various fields including social and physical sciences, engineering and economy.

A model associating a response (dependent) variable with one or more explanatory (independent) variables with function linear in parameters is called linear regression model. The form of a linear regression model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{2.1}$$

where y_i is the response variable, x_i's are the explanatory variables, n is the number of observation, m is the number of independent variables and ε_i is the error component. β parameters which represent the coefficients attributed to explanatory variables to predict response variable, is usually calculated by means of the least squares estimation (LSE) technique. This technique minimizes the sum of squares of differences between the observed and calculated response variables.

Linear regression analysis assumes the error term to distribute normal with a mean, $E(\varepsilon_i)$, equal to zero. So,

$$\mu_i = E(y_i|x_i) = E(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im},$$
$$i = 1, 2, \ldots, n \tag{2.2}$$

Furthermore, the variance of the error terms, σ^2, is assumed to be constant, for different levels of response variables and independent from each other. This also means that;

$$E(y_i|x_i) \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}, \sigma^2) \tag{2.3}$$

Besides, the lack of correlation between error terms leads the response variable to be also uncorrelated.

2.2. Generalized Linear Models

The concept of GLM refers to a broader category of modelling. It is an extension of linear models in terms of the distribution of the response variable. It provides solutions to the cases where the distribution of the response variable is from exponential family of distributions. A distribution is considered as a member of exponential family of distributions if its density function, $f(.)$, is in the following form:

$$f(y_i, \Theta_i, \sigma) = \exp\{b(\Theta_i, \sigma)y_i + c(\Theta_i, \sigma) + d(y_i, \sigma)\} \tag{2.4}$$

where Θ_i is the location parameter and σ is the nuisance parameter.

Generalized Linear Models apply a function to describe how the expected value of the response variable depends to the explanatory variables or linear predictor, called a link function:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}, \quad i = 1, 2, \ldots, n. \qquad (2.5)$$

So, a linear relation can be built by means of a link function in these cases while the range of transformed mean varies between $-\infty$ and $+\infty$ in line with the Normal distribution.

Link function differs according to the distribution of the response variable. For instance, when the distribution of response variables is Exponential or Gamma, the link function $g(\mu_i)$ equals to $-\mu_i^{-1}$. Also, it equals to μ_i^{-2} and $\ln(\mu_i)$ for Inverse Gaussian distribution and Poisson distribution; respectively.

Link function, for the cases where the response variable distributes Normal, is identity link in which $g(\mu_i)$ is equal to μ_i. It indicates that an ordinary linear regression model can be considered as a subset or a special case under Generalized Linear Models.

The estimation of parameters in GLM is conducted by means of Maximum Likelihood Estimation (MLE) with iterative methods like Newton-Raphson Method and Iterative Reweighted Least Squares Method.

2.3. Logistic Regression Models

There are many circumstances where the response variable is not continuous but binary instead. In such cases where there are two certain options for the response variable, like registered to social security institutions or not, survival or death, cancer

or not, present or absent, logistic regression analysis is conducted to observe its relationship with explanatory variables. Namely, Logistic Regression Model, which is also known as Logit Model, is a form of Generalized Linear Model to be applied when the response variable distributes Bernoulli and have the following pmf:

$$P(y_i) = \begin{cases} 1 - p_i & \text{where } y_i = 0 \\ p_i & \text{where } y_i = 1 \end{cases} \tag{2.6}$$

So, the conditional mean (or the expected value of the response variable), $E(y_i|x_i)$, is equivalent to the probability of the interested event's occurrence. Namely,

$$E(y_i|x_i) = \mu_i = [(1 - p_i) * 0] + [p_i * 1] = p_i \tag{2.7}$$

and

$$1 - E(y_i|x_i) = 1 - \mu_i = P(y_i = 0) = 1 - p_i. \tag{2.8}$$

Since $E(y_i|x_i)$ is, at the same time, a probability in logistic regression case;

$$0 \leq E(y_i|x_i) = \mu_i \leq 1. \tag{2.9}$$

In these premises, logistic regression function is designed to keep the expected value within the range of 0 and 1 as follows:

$$E(y_i|x_i) = \mu_i = \frac{\exp\{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}\}}{1+\exp\{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}\}} = \frac{1}{1+\exp\{-(\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im})\}} \quad i = 1, 2, \ldots, n \tag{2.10}$$

On the other hand, the link function can be derived from the logistic regression function by taking its inverse as:

$$g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}, \quad i = 1, 2, \ldots, n \tag{2.11}$$

This function is also called "logit transformation function". Since the graph of $E(y_i|x_i)$ and X is S-Shaped as in Figure 2.1 and the range of $E(y_i|x_i)$ is between 0 and 1, the logit function, or also the logistic regression function, is derived from Logistic Distribution.

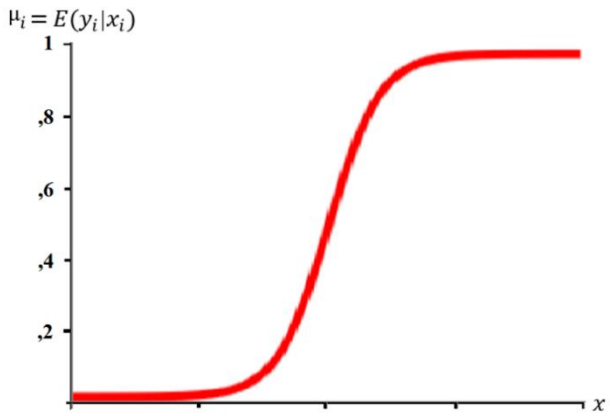

Figure 2.1. Logistic Regression Function

2.3.1. Interpretation of Logistic Regression Models

The interpretation of coefficients in a fitted logistic regression model is not as straightforward as linear regression model depending on the dichotomous property of the response variable. Namely, it is not possible to apply well known "one unit change in the independent variable leads β unit change in the dependent variable" interpretation.

Considering the definition of odds as the probability of an event's occurring divided by the probability of an event's not occurring, the linear model provided by the logit function can be considered as the log odds of event's occurring, since the logit function $g(\mu_i) = \ln\left(\frac{p_i}{1-p_i}\right)$ where $p_i = \Pr(y_i = 1)$ and $1 - p_i = \Pr(y_i = 0)$.
Accordingly, it can be inferred that "one unit change in one independent variable leads β unit change in the log odds of occurring".

A clearer interpretation can be made in terms of odds ratio (OR) which is the ratio of odds of occurring for an independent variable's two different values (or its presence and absence). In order to reach odds ratio, we need to exponentiate the natural logarithms of odds of occurring with certain values of independent variables;

$$[\log.\mathrm{odds}(x, x_{im})] = [\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m(x_{im})] \tag{2.12}$$

$$\mathrm{Exp}\{\log.\mathrm{odds}(x)\} = \mathrm{odds}(x) = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}\} \tag{2.13}$$

$$\mathrm{odds}(x_i) = \exp\{\beta_0\} * \exp\{\beta_1 x_{i1}\} * \ldots * \exp\{\beta_m(x_{im})\} \tag{2.14}$$

In case of λ unit increase in one of the independent variables, for instance in x_{im}, then

$$\mathrm{odds}(x_i, x_{im} + \lambda) = \exp\{\beta_0\} * \exp\{\beta_1 x_{i1}\} * \ldots * \exp\{\beta_m (x_{im} + \lambda)\} \quad (2.15)$$

$$= \exp\{\beta_0\} * \exp\{\beta_1 x_{i1}\} * \ldots * \exp\{\beta_m x_{im}\} * \exp\{\beta_m \lambda\} \quad (2.16)$$

$$\frac{\mathrm{odds}(x_i, x_{im} + \lambda)}{\mathrm{odds}(x_i)} = \frac{\exp\{\beta_0\} * \exp\{\beta_1 x_{i1}\} * \ldots * \exp\{\beta_m x_{im}\} * \exp\{\beta_m \lambda\}}{\exp\{\beta_0\} * \exp\{\beta_1 x_{i1}\} * \ldots * \exp\{\beta_m (x_{im})\}} = \exp\{\beta_m \lambda\} \quad (2.17)$$

Accordingly, we can conclude the interpretation as "λ unit of increase in x_m leads $exp\{\beta_m \lambda\}$ unit change in the odds of occurring".

Predicted probabilities in a logistic regression can also be used for the interpretation (Long, 1997). The researcher can focus on the probabilities for different levels of independent variables. The change in the predicted probability can be clearly illustrated with a graph while the interested independent variable varies in its range with all else independent variables set constant. If the minimum and the maximum values of the predicted probabilities attributed is between 0.2 and 0.7; then the relationship between the interested independent and the dependent variable can be considered as linear due to the S-shaped logistic regression function, and interpreted accordingly. Also, tabulated values of probabilities at selected levels of independent variables may help to interpret the regression. The levels of the independent variables other than the interested one can be determined as their means or modes or in accordance with the research questions.

Using marginal effects for the interpretation is another way in this context. Marginal effect indicates the effect on $E(y_i|x_i)$ of a change in an independent variable. Namely, it describes the average effect of changes in explanatory variables on the change in the probability of outcomes. For logistic regression, marginal effect provides the simplicity for expressing the effect of an independent variable on P(Y=1). In other

words, it figures the change in the probability of occurrence with the change in the independent variable. Especially in the models with categorical covariates, it is a straightforward method of interpretation.

In order to calculate the marginal effect of an independent variable, other independent variables than the variable in question is controlled. However, the approach to control the independent variables other than the one in question differs. For instance, Marginal Effect at Representative Values (MER) is calculated when the other independent variables are assumed to be at their representative values (mean, mode, etc.). On the other hand, Average Marginal Effect (AME) is calculated by taking the average of the changes in probability for each observation. The preference of the type of the marginal effect mainly depends on the research question.

2.3.2. Model Fitting in Logistic Regression via Maximum Likelihood Estimation

Parameter estimation in Logistic Regression Model, as a version of Generalized Linear Model, is conducted by means of Maximum Likelihood Method which provides estimation of the parameters that maximize the likelihood function (Hosmer et al, 2013). Since the response variable distribute Bernoulli, the pdf of sample observation is

$$f_i(y_i) = \mu_i^{y_i}.(1-\mu_i)^{1-y_i}, \quad i = 1, 2, \dots, n. \tag{2.18}$$

So the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} \mu_i^{y_i}.(1-\mu_i)^{1-y_i} = \prod_{i=1}^{n} \frac{\mu_i^{y_i}.(1-\mu_i)}{(1-\mu_i)^{y_i}} \tag{2.19}$$

with the log-likelihood function

$$\ln[L(\beta)] = \sum_{i=1}^{n} y_i . \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \sum_{i=1}^{n} \ln(1-\mu_i). \tag{2.20}$$

Since

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} \tag{2.21}$$

and

$$1 - \mu_i = [1 + \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}\}]^{-1}, \tag{2.22}$$

the log-likelihood function becomes

$$\ln[L(\beta)] = \sum_{i=1}^{n} y_i . (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}) - \sum_{i=1}^{n} \ln(1 + \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}\}). \tag{2.23}$$

In order to find maximum likelihood estimates of the coefficients, we are supposed to take the derivative of the log-likelihood function with respect to each parameter and set equal to zero; i.e.

$$\frac{\partial \ln[L(\beta)]}{\partial \beta_k} = \sum_{i=1}^{n} x_{ik}(y_i - \mu_i) = 0, \quad k = 1, 2, \ldots, m. \tag{2.24}$$

Since the equations are not linear in terms of β parameters, iterative reweighted least squares method can be applied in order to reach maximum likelihood estimates of the regression. Luckily, there are many statistical software programs capable to conduct the iteration procedure and solve the equations.

2.3.3. Significance Testing in Logistic Regression

The idea behind the significance testing in logistic regression is similar to the one in linear regression: Comparing the extent of models' informativeness with and without one or more variables in question. Regarding the logistic regression, this comparison can be conducted via likelihood ratio test statistic to observe the significance of one or more coefficients or the whole model at the same time. This statistic tests the constraints on a model as follows:

$$\mathrm{LR} = \frac{\mathrm{L}(\mathrm{M}_{\mathrm{Constrained}})}{\mathrm{L}(\mathrm{M}_{\mathrm{Unconstrained}})} = \frac{\text{likelihood without the variable}}{\text{likelihood with the variable}} \tag{2.25}$$

where $M_{Constrained}$ is a model with "k" parameters and $M_{Unconstrained}$ is a model with "m" parameters, k < m. Also, it should be noted that $M_{Constrained}$ is formed from $M_{Unconstrained}$ by applying the constraint equalizing "m-k" number of β's in question to 0.

Since maximum likelihood estimators, as in the logistic regression, asymptotically distribute normal, -2ln(LR) approximates to distribute Chi-squared with degrees of freedom equal to m-k as $n \to \infty$. This provides "G" which is the likelihood ratio test statistic to service as a tool for hypothesis testing in logistic regression analysis:

$$\mathrm{G} = -2\ln(\mathrm{LR})$$

$$= -2[\text{log. likelihood without the var.} - \text{log. likelihood with the var.}]. \tag{2.26}$$

The asymptotic normal distribution of maximum likelihood estimators also enables a method to be applied for hypothesis testing of individual coefficient: Wald Statistic as

$$W_J = \frac{\widehat{\beta_J}}{\widehat{SE}(\widehat{\beta_J})} \tag{2.27}$$

where W_J asymptotically distributes normal, $\widehat{\beta_J}$ is the maximum likelihood estimator of β_J coefficient in question and $\widehat{SE}(\widehat{\beta_J})$ is the standard error of the $\widehat{\beta_J}$.

Most statistical software, including R, give individual parameter significance in terms of Wald statistic because there is not a need to compute an additional fit in line with the null hypothesis (Tutz, 2012).

Wald statistic can also be used to test the significance of multiple parameters. In this case, the formulation can be formed in terms of vectors and matrices as follows:

$$W = \widehat{\beta_r}' \, [\widehat{Var}(\hat{\beta}_r)]^{-1} \widehat{\beta_r} \tag{2.28}$$

where W distributes Chi-squared with degrees of freedom equal to the number of constrained coefficients in line with the null hypothesis, $\widehat{\beta_r}$ is the vector of constrained coefficients which is a sub-vector of parameter vector $\hat{\beta}$ and $\widehat{Var}(\hat{\beta}_r)$ is the corresponding submatrix of the varcov matrix of $\hat{\beta}$. Furthermore, revised versions of Wald test statistic is utilized to test the significance of various functions of the coefficients (Hosmer et al, 2013).

2.3.4. Goodness of Fit in Logistic Regression

There are various measures in order to determine if a logistic regression fits the observations while all important explanatory variables are taken into consideration with their correct functional forms. Pearson Chi-Squared Test, Deviance Test, Hosmer-Lemeshow Test, Pseudo R^2 Test, Classification Tables and Area Under

Receiver Operating Characteristic (ROC) Curve are among the most frequently used tools for this purpose.

2.3.4.1. Pearson Chi-squared Test

In Pearson Chi-squared Test, after establishing a logit model based on "n" observations with a binary dependent variable and categorical independent variables, a table of all possible combinations of independent variables which is equal to "P" is drawn. The table indicates the corresponding observed counts for every combination of independent variables and each category of the binary dependent variable. Afterwards, expected counts for each combination are calculated according to the model (Tang et al, 2012). The layout of the table will be as follows:

Independent Variable Pattern	**Observed Count (Expected Count)**	
	y=1	**y=0**
X_1	n_{11} (E_{11})	n_{12} (E_{12})
X_2	n_{21} (E_{21})	n_{22} (E_{22})
:	:	:
X_P	n_{P1} (E_{P1})	n_{P2} (E_{P2})

Pearson Chi-squared Test is based on comparison of observed and expected counts as follows:

$$X^2_{\text{Pearson}} = \sum_{j=1}^{P} \sum_{k=1}^{2} \frac{(n_{jk} - E_{jk})^2}{E_{jk}} \tag{2.29}$$

where $1 \leq j \leq P$ and $k = 1$ or 2. $X^2_{Pearson}$ test statistic asymptotically distributes chi-squared with degrees of freedom equal to $P-\ell$ where ℓ is the number of parameters to be estimated by the model.

2.3.4.2. Deviance Test

Another test procedure used to compare the observed number and the expected number of event is the Deviance Test. It utilizes the same independent variable pattern as in Pearson Chi-squared Test with a different formulation:

$$X^2_{\text{Deviance}} = 2\sum_{j=1}^{P}\sum_{k=1}^{2} n_{jk} \ln\left[\frac{n_{jk}}{E_{jk}}\right]. \tag{2.30}$$

$X^2_{Deviance}$ test statistic distributes asymptotically chi-squared with degrees of freedom equal to P–ℓ. Deviance test is a likelihood ratio test of the fitted model and the saturated model which is hypothetically complex and results as a perfect fit.

In both Pearson Chi-squared Test and Deviance test, the distribution of the test statistic departs from Chi-square Distribution in case of presence of continuous independent variable(s) unless it is grouped. In other words, they mislead if the expected number of events or non-events in each cell are less than 5 (Allison, 2013).

2.3.4.3. The Hosmer-Lemeshow Test

Considering the problems about satisfying minimum 5 in each cell condition in Pearson Chi-squared Test and Deviance Test, Hosmer and Lemeshow (1980) proposed a method grouping the combinations according to their predicted values and changing the pattern table above by setting up groups with approximately equal sizes.

The Hosmer-Lemeshow Test initially puts the subjects in order according to their predicted probabilities and separates them into (usually) 10 groups with almost equal sizes. Afterwards, expected counts for each cell are calculated from the fitted model similar to Pearson Chi-squared Test and Deviance Test and the test statistic is calculated as follows:

$$X_{H-L}^2 = \sum_{j=1}^{g} \sum_{k=1}^{2} \frac{(n_{jk}-E_{jk})^2}{E_{jk}} \tag{2.31}$$

where g is the number of groups. X_{H-L}^2 test statistic distributes approximately chi-squared with degrees of freedom equal to g-2 (Tang et al, 2012).

2.3.4.4. Classification Tables

Observed values and predicted values of the dependent variable are summarized in a classification table in this method. Predicted values of the dependent variable are derived from the predicted probabilities calculated from the fitted model. A cut-off value is defined in order to binarize the predicted values of the dependent variable. Thus, the layout of the classification table (also known as confusion matrix) is as follows:

		Observation		
		Success	Fail	Total
Classification	Success	True Positive	False Positive	Total Positive Predicted
	Fail	False Negative	True Negative	Total Negative Predicted
		Total Positive Observed	Total Negative Observed	Total

where the main diagonal of the table includes the true classifications and the substitute diagonal includes the false classifications. Derived from the table, the proportion of true classifications over observations $\left(\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}\right)$ gives us the accuracy rate as a measure of goodness of fit (Hosmer et al, 2013).

2.3.4.5. Area Under ROC Curve

The method of Area Under ROC Curve also benefits from the classification table, however, it summarizes all of the tables for each cut-off points. This method is based on "sensitivity" and "specificity". Namely, sensitivity is the proportion of true classification of success observations$\left(\frac{TP}{TP+FN}\right)$ and specificity is the proportion of true classification of failure observations$\left(\frac{TN}{TN+FP}\right)$. ROC Curve illustrates sensitivity and 1-specificity for each cut-off point as follows:

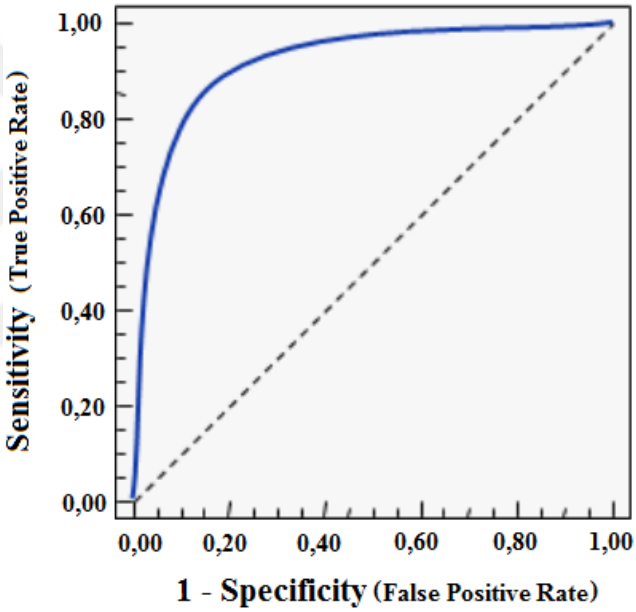

Figure 2.2. Area Under ROC Curve

As can be seen in Figure 2.2, the curve starts from the origin where sensitivity is equal to 0 and specificity is equal to 1 which means the cut-off point is assumed to be equal to 1. On the other hand, the curve ends at where sensitivity is equal to 1 and specificity is equal to 0 which means the cut-off point is assumed to be equal to 0. Naturally, a ROC Curve with higher sensitivity and specificity has a shape with higher top-left

corner which indicates a larger area under the curve (Hosmer et al, 2013). The interpretation of Area under ROC Curve is as follows:

ROC=0.5	no discrimination
0.5<ROC<0.7	poor discrimination
0.7≤ROC<0.8	acceptable discrimination
0.8≤ROC<0.9	excellent discrimination
ROC≥0.9	outstanding discrimination

2.3.4.6. Pseudo $\mathbf{R^2}$ Test

There are various methods to apply $\mathbf{R^2}$ to Logistic Regression case. $\boldsymbol{Pseudo}\ \mathbf{R^2}$, developed by McFadden (1974) is the most used one among all. Analogous to the Linear Regression case, this method compares the best fit with the worst fit by means of log-likelihoods instead of the residuals as follows:

$$\mathbf{Pseudo\ R^2} = 1 - \frac{\ln\,[\hat{\mathrm{L}}(\mathrm{M_{fitted}})]}{\ln\,[\hat{\mathrm{L}}(\mathrm{M_{intercept}})]} \tag{2.32}$$

where M_{fitted} is the model to be determined on its goodness of fit, $\mathrm{M_{intercept}}$ is the model with only intercept and ln $[\hat{\mathrm{L}}(\mathrm{M_{intercept}})]$ is equal to ln[$\hat{\mathrm{L}}$(Overall Probability of Occurrence)] (Long, 1997).

CHAPTER 3

GROUP COMPARISON PROBLEM IN LOGISTIC REGRESSION

This chapter focuses on the group comparison problem in Logistic Regression Models. In order to reveal the problem more clearly, latent propensity perspective to the Logistic Regression Modelling is emphasized and the relevant discussions in the literature are given.

3.1. Latent Propensity Perspective in Logistic Regression

The dependent variable in logistic regression is binary and observed as 0 or 1. For instance, an observed person is registered to social security institutions or not. However, some researchers find this "one size fits all" perspective strict since it ignores this person's extent of closeness to work unregistered. The ones observed as registered are not all in the same risk of working unregistered. Namely, there is a latent propensity to work unregistered for each person and this is indicated as y^*. On the other hand, this perspective is useful to understand the effect of the logistic distribution in the field of logistic regression analysis.

y^* is a hypothetical value which cannot be observed in reality and ranges between $-\infty$ and $+\infty$. The relationship between the latent propensity and the observed independent variables can be shown with a linear model as follows:

$$y^* = \beta_0^* + \beta_1^* x_{i1} + \cdots + \beta_m^* x_{im} + \sigma\varepsilon_i \tag{3.1}$$

where $\beta^{*\prime}$s are latent propensity model coefficients and σ is the adjustment parameter (factor) for the variance of the error term, not the variance of the error term itself.

The change in the latent propensity leads to a change in the observed y after a certain point as follows:

$$y_i = \begin{cases} 1 & \text{if } y^* > \delta \\ 0 & \text{if } y^* \leq \delta \end{cases} \tag{3.2}$$

where δ is the threshold and usually equal to 0.

Since $P(Y = 1) = P(y^* > 0)$,

$$P(Y = 1) = P(\beta_0^* + \beta_1^* x_{i1} + \cdots + \beta_m^* x_{im} + \sigma\varepsilon_i > 0)$$

$$= P\left(\frac{\beta_0^*}{\sigma} + \frac{\beta_1^*}{\sigma} x_{i1} + \cdots + \frac{\beta_m^*}{\sigma} x_{im} + \varepsilon_i > 0\right) = P(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i > 0). \tag{3.3}$$

Realize that all β's and x's in Equation (3.3) are fixed, and the error term is the only random variables. However, the distribution of the error term is unknown because y^* is unobserved. In order to be able to apply Maximum Likelihood Estimation, the distribution of the error term is assumed to be Logistic with variance equal to $\pi^2/3$ and $E(\varepsilon)$ equal to 0. Thus,

$$P(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i > 0) = P(\varepsilon_i > -(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}). \tag{3.4}$$

The probability in Equation (3.4) can be calculated from the cumulative distribution function (cdf) of the random variable, ε_i in this case. Implementing the assumption of logistic distribution for the error terms,

$$P(y = 1) = \frac{1}{1+\exp\{-(\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im})\}} \tag{3.5}$$

which is equal to the logistic regression function (Long, 1997; Buis, 2016).

3.2. Group Comparison Problem in Logistic Regression

The difference of an independent variable's level of effect to the change in the response for different populations is an important question and a source for inference for a researcher. Accordingly, Chow (1960) started a new thread of testing the equality of coefficients across groups limited to linear regressions.

In OLS, Wald Chi-squared Statistic is the most common method in this context, which is calculated as follows:

$$\frac{[\hat{\beta}_k^{(1)}-\hat{\beta}_k^{(2)}]^2}{[\widehat{SE}(\hat{\beta}_k^{(1)})]^2+[\widehat{SE}(\hat{\beta}_k^{(2)})]^2} \tag{3.6}$$

in order to test the hypothesis H_0: $\hat{\beta}_k^{(1)} = \hat{\beta}_k^{(2)}$ where $\hat{\beta}_k^{(1)}$ and $\hat{\beta}_k^{(2)}$ are the estimated coefficients for the models of Group 1 and 2; respectively, while $\widehat{SE}\left(\hat{\beta}_k^{(1)}\right)$ and $\widehat{SE}\left(\hat{\beta}_k^{(2)}\right)$ are the standard errors of the corresponding estimated coefficients (Allison, 1999).

Applying the methods to test the difference of coefficients across groups in linear regression to logistic regression is handicapped because the difference between coefficients might be a result of unobserved heterogeneity (Allison, 1999; Williams, 2009; Long, 2009; Mood, 2010). This can be indicated by means of latent propensity perspective of logistic regression. However, unobserved heterogeneity can be summarized as the variation in the response variable arise from non-observed or omitted covariates which actually affects the response variable.

Since $\hat{\beta}_k = \frac{\hat{\beta}_k^*}{\sigma}$ comparison of $\hat{\beta}_k^{(1)}$ and $\hat{\beta}_k^{(2)}$ derived from logistic regression with the common used methods in OLS is incorrect considering that σ's may change for Groups 1 and 2 (Amemiya,1985; Maddala, 1983). Accordingly, there is a scalar identification problem of logistic regression coefficients led by the residual variation (Allison, 1999). So, logit coefficients may be confounded by the residual variation and the difference between the coefficients may be artificial. In other words, apparent differences in coefficients may not mean a true difference stemming from causal effects. Thus, a significant difference between coefficients across groups may be insignificant and vice versa.

Allison (1999) proposed a method to overcome this problem by eliminating the differences in error variances of models for two populations. In order to do so, an assumption depending on identical $\hat{\beta}^*$ coefficients for some variables across groups was implemented and the ratio of the corresponding $\hat{\beta}$'s gave a ratio of σ factors for each population. This ratio was used as a tool to make the coefficients comparable and the likelihood ratio methodology were implemented in order to test the hypothesis of equality.

Williams (2009) rejected Allison's claim to routinely impose this method and argued that Allison's model was a heteroscedastic logistic regression model and provided biased estimates when the difference of residual variability is low. Furthermore, he implemented Allison's method to test models constructed from simulated data for two different groups with no difference in residual variability and found that Allison's

method indicated a difference. He also concluded that the heteroscedastic logistic regression model which was the case in Allison's study, was a special case of heterogeneous choice model which is also known as location-scale model and using the heterogeneous choice model helps the determinants of the heteroscedasticity to be modelled in a more satisfactory way for cases both with binary and ordinal dependent variables. He also proposed a method providing a solution based on identifying the degree of heterogeneity and intervening accordingly similar to Allison in this aspect.

Mood (2010) also underlined the problem of unobserved heterogeneity in logistic regression and concluded its problematic results with a more holistic approach by including the problems about the interpretation of odds ratios and the comparison of odds ratios across models with different independent variables in addition to comparison of odds ratios across models with same independent variables but different groups. Winship & Mare (1984) and Karlson et al. (2012) focused on y-standardization in order to compare coefficients across nested models with the same samples.

Moods (2010) recommended to study on continuous dependent variables as much as possible considering the problems special to logistic regression. She then discussed some solutions to the problems and categorized the solutions as the ones based on odds ratios and the ones based on probability changes. She concluded the different effect estimates including Allison's and Williams's procedures, average marginal effect, average partial effect and linear probability model methods as eligible for comparison of coefficients across groups.

Contrary to Mood (2010), Buis (2016) argued that unobserved heterogeneity is not a problem in logistic regression because the dependent variable is not the latent variable. Rather, he focused on the cases where unobserved heterogeneity is a real influence on the coefficients and desirable in this sense.

Similar to Buis (2016), Kuha and Mills (2017) argued that coefficients of logistic regression analysis are eligible to compare across groups if the main research question

is not based on the latent propensity (y^*) and they tell us that most of the studies in social sciences are not really interested in the latent propensity.

Angrist (2001), Ai & Norton (2003) recommended focusing on the probabilities rather than the latent variable for the interpretation of limited dependent variable models including logistic regression models. Similarly, Long (2009) found the test of predicted probabilities as a more convenient alternative to the tests based on latent propensity. He criticized Allison's test since it is based on the assumption of equal coefficients for at least one independent variables of each group. He evaluated such an assumption as weak since in most cases researchers do not have scientific evidence indicating that kind of equality. Rather, Long (2009) argued that test of probabilities based on the models is not affected by unobserved heterogeneity and justifies his argument as follows:

$$P(Y = 1) = P(y^* > 0) = P(\beta_0^* + \beta_1^* x_{i1} + \cdots + \beta_m^* x_{im} + \sigma\varepsilon_i > 0)$$

$$= P\left(\frac{\beta_0^*}{\sigma} + \frac{\beta_1^*}{\sigma} x_{i1} + \cdots + \frac{\beta_m^*}{\sigma} x_{im} + \varepsilon_i > 0\right) = P(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i > 0)$$

$$= P(\varepsilon_i < \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}) = F(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m) \qquad (3.7)$$

where F is the cdf of logistic distribution for logistic regression and the predicted probability can be calculated by means of the cdf. Considering the scalar identification of logit coefficients, $\hat{\beta}_k = \frac{\hat{\beta}_k^*}{\sigma}$, $P(\varepsilon_i < \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im})$ in Equation (3.7) is multiplied by a constant "Λ" in order to simulate the effect of scalar identification to the predicted probability.

Since

$$P(\Lambda\varepsilon_i < \Lambda\beta_0 + (\Lambda\beta_1)x_{i1} + \cdots + (\Lambda\beta_m)x_{im}) = P(\varepsilon_i < \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}), \tag{3.8}$$

it can be concluded that the predicted probability is not affected by the scalar identification. Accordingly, Long (2009) proposes a method for the test of predicted probabilities based on delta method which is used to compute the variance of functions of ML estimates such as predicted probabilities as follows:

$$\mathrm{var}[\mu(x_i)] = f(x_i'\beta).x_i'.\mathrm{var}(\hat{\beta}).x_i.f(x_i'\beta) \tag{3.9}$$

where μ is the predicted probability, f is the pdf of the standardized logistic distribution, $\hat{\beta}$ is the vector of ML estimates and $var(\hat{\beta})$ is the variance-covariance matrix of the estimated parameters. So, the variance of the predicted probability difference is

$$\mathrm{var}[\mu(x_i)_{Group1} - \mu(x_i)_{Group2}] = \mathrm{var}[\mu(x_i)]_{Group1} + \mathrm{var}[\mu(x_i)]_{Group2} \tag{3.10}$$

and the Z-statistic with asymptotic normal distribution to test $H_0: \mu(x_i)_{Group1} = \mu(x_i)_{Group2}$ is

$$Z = \frac{\mu(x_i)_{Group1} - \mu(x_i)_{Group2}}{\sqrt{\mathrm{var}[\mu(x_i)_{Group1} - \mu(x_i)_{Group2}]}} \tag{3.11}.$$

Also, Long's (2009) method made the difference of predicted probabilities available to construct a confidence interval. Furthermore, it enables the researcher to indicate the difference between the groups along the different levels of the same independent variable.

Long & Mustillo (2018) enlarged the method of using predicted probabilities for comparison purposes among the groups by marginal effects to probabilities: discrete change at representative values where marginal effect is estimated at representative values of X's and average discrete change where the average marginal effect is estimated conditional on the observed values of X's for each observation.

In this study, Long's methods to evaluate the group differences based on the predicted probabilities and the average marginal effect were implemented since the probabilities and the marginal effects are not scaled by unobserved heterogeneity and more straightforward to interpret. Within this context, depending on the fact that probabilities can be calculated by means of all independent variables jointly, independent variables which are not subject to the primary interest are determined at their representative values in line with the researcher's questions.

CHAPTER 4

UNREGISTERED EMPLOYMENT

This chapter briefly defines the concept of unregistered employment and reveals the reasons and the impacts of the phenomenon as well as fighting methods with it. Afterwards, unregistered employment particularly in Turkey is analyzed taking its characteristics into consideration.

4.1. Informal Sector

International Labour Organization (ILO) introduced the concept of the "informal sector" in a comprehensive report published in January, 1972 which analyzed the employment market problems in Kenya. This report characterized the informal sector as including economical activities which are small in scale, limited by simple technologies, limited amount of capital and lack of links with the formal sector (ILO,1972).

Even though the informal sector was evaluated as temporary at the very beginning; it was accepted to have a permanent characteristic indeed afterwards (Charmes, 1990). Considering the informal sector as a provider of employment and income, there was a dilemma for ILO and the governments between promoting the sector or seeking to extend regulation and social protection. The report of the 78th Session of the International Labour Conference (ILO, 1991) underlined that "there can be no question of the ILO helping to 'promote' or 'develop' an informal sector as a convenient, low-cost way of creating employment unless there is at the same time an equal determination to eliminate progressively the worst aspects of exploitation and inhuman working conditions in the sector".

In the same report (ILO, 1991), a clearer definition of informal sector is made as "very small-scale units producing and distributing goods and services, and consisting largely of independent, self-employed producers in urban areas of developing countries, some of whom also employ family labour and/or a few hired workers or apprentices; which operate with very little capital, or none at all; which utilize a low level of technology and skills; which therefore operate at a low level of productivity; and which generally provide very low and irregular incomes and highly unstable employment to those who work in it."

4.2. Unregistered Employment

According to the 22nd Article of the Universal Declaration of Human Rights, "everyone, as a member of society, has the right to social security and is entitled to realization, through national effort and international co-operation and in accordance with the organization and resources of each State, of the economic, social and cultural rights indispensable for his dignity and the free development of his personality." In this sense, States are responsible to establish and maintain a social security system.

ILO Convention No: 102 defines the social security system as a system which covers the benefits in cases of old-age, invalidity, survivors, sickness, maternity, employment injury, unemployment, medical care and family benefits. Considering its function and scope, the financing of this system is critical while unregistered employment can be accounted as one of the main reasons in the deficit of the social security system. In addition to its damage to the system itself, this phenomenon leaves the individuals unprotected towards socials risks.

ILO defines the informal (unregistered) employment as "the total number of informal jobs, whether carried out in formal or informal enterprises, or the total number of persons engaged in informal jobs during a given reference period" (ILO, 2002). Accordingly, we can evaluate unregistered employment as a phenomenon taking place not only in informal sector but also in formal sector.

On the other hand, European Union defines the undeclared work as "any paid activities that are lawful as regards their nature but not declared to public authorities, taking into account differences in the regulatory system of Member States" (European Commission, 2007).

There are various synonyms of unregistered employment depending on differing approaches to the concept. It is often referred as informal employment, black labour, hidden employment or undeclared work. The concept of “informal employment” are more frequently used in applied researches, wherein it is defined as the number of people working in the informal labour market, as the illegal purchase and sale of labour force without an employment contract and ignoring laws that regulate labour relations (ILO, 2013).

Considering the main indicators of unregistered employment as not declaring or partially declaring to authorities, unregistered employment can be classified into 3 groups as follows:

- Fully undeclared labour contract to Social Security Institution,
- Partially undeclared in terms of income dependent on social security contribution,
- Partially undeclared in terms of working days.

According to ILO (2018) unregistered employment rate is 61.2% in the World, while it is 18.3% in developed countries and 69.6% in emerging and developing countries. Vast majority of employment in Africa (85.8%) is unregistered. On the other hand, unregistered employment is almost equal to each other in Arab States with Asia and the Pacific (68.6% and 68.2% respectively). Furthermore, in Americas including Latin, Central, South and Northern America, four in every ten person work unregistered while one in every four persons is working unregistered in Europe and Central Asia.

4.2.1. Reasons of Unregistered Employment

The reasons of unregistered employment can be classified as macroeconomic reasons, social reasons and labour market status. However, it should be noted that the subject classes are in interaction with each other.

4.2.1.1. Economic Reasons

Unregistered employment is a chronic phenomenon in countries with relatively high unemployment and inflation rates. These indicators directly affect the individuals' disposition to work without declaring to the relevant authorities.

Individuals at higher risk of unemployment may accept to waive their right of social security (Frey and Weck-Hanneman, 1984). Also, employers may tend to dismiss the employees not accepting to work without social security with the opinion to easily find another employee to work under this condition. Similarly, inflation leads the deprivation of households' income levels while forcing the breadwinners to find new financial sources, and the employer to reduce the employment costs.

On the other hand, the share of sectors in the economy is highly determinant for the rate of unregistered employment in a country. Unregistered employment is very common in sectors which are based on the employment of family members and which are difficult to inspect in terms of declaration like agriculture and construction. Therefore, economies primarily built on these sectors tend to employ undeclared higher. In addition, economies heavily comprised of small and middle enterprises suffer from the same problem depending on this kind of enterprises' lack of institutionalism and inspectability (Güloğlu, 2005; DPT, 2001, Sarılı, 2004).

4.2.1.2. Social Reasons

Income inequality, poverty, immigration and population increase can be considered under social reasons of unregistered employment. These elements are interrelated with each other as well.

Poor groups with a low share of income with additional job demand or women (and even children) who cannot participate in the labour market in terms of their qualifications seek employment in the informal sector. Moreover, the increasing number of people living in cities and the decreasing possibility of living in better conditions as a result of the migration and population increase lead to the growth and diversification of the informal sector (Açıkalın, 2007: 48).

4.2.1.3. Reasons Arising from Labour Market Status

In addition to social and economic reasons, the current situation and trends in the labour market also lead to the unregistered employment. Depending on the high financial burden of labour force including the premiums and the taxes, employers may search for alternative ways to avoid costs and prefer not to declare the employment to the authorities where the inspection mechanism does not work well and the level of consciousness is not high.

The transition from the traditional form of employment to the flexible employment practices such as part-time work, temporary work and home based work in order to reduce the costs of enterprises and increase their competitiveness lead to the lack of social security of the employees and the unregistered employment (Yavuz, 1995; Van Eyck, 2003). On the other hand, subcontracting system, which is highly associated with unregistered employment, becomes widespread due to the fact that it is seen as a cost reduction method by going beyond its definition (Van Eyck, 2003).

4.2.2. Impacts of the Unregistered Employment

Unregistered employment has diverse negative consequences on many sides of the labour market. The challenges posed by the unregistered employment can be categorized under three main topics including the financial impacts, economic impacts and impacts of employees.

4.2.2.1. Financial Impacts

The primary financial effect of the unregistered employment is the loss of social security premium of the state. Social security system loses revenue depending on the unregistered employees; on the other hand, the conditions to benefit from the rights and services of the system are facilitated for different political reasons, thus increasing the need for additional financing.

In the case of unregistered employment, in addition to the loss of social security premiums, there is also tax loss. In other words, the tax burden of the unregistered employment is shifted to the formal sectors (Schneider, 2000).

4.2.2.2. Economic Impacts

Measuring unregistered employment completely is difficult by its nature. Thereby, informality strongly manipulates the macroeconomic data to be used by the policy makers (Çetintaş and Vergil, 2003). Also, it provides unfair competition advantage to some companies. This advantage causes the extension of unregistered employment to the other companies in the economy. Besides, it is associated with low wages and poor working conditions, thus creating an inefficient economy.

Another effect of informal employment on the economy is the creation of employment. At the same time, especially in developing countries, as a result of unregistered employment, the taxes and premiums remaining in the employer have the

ability to create additional value and increase the growth and employment if they are directed to production (Adam and Ginsburgh, 1985). However, it should not be overlooked that this effect, which is seen as positive at first glance, causes acquiescence of unregistered employment in the society and it becomes more widespread during the times of recession.

4.2.2.3. Impacts on Employees

In case of informality, an employee is open to social risks mainly including occupational diseases and accidents, unemployment, maternity, old age, invalidity/disability and death. Also, those workers whose employment is not declared to the authorities are deprived of the right to strike and trade union rights. Even worse, child labour is an inherent characteristic of unregistered employment.

4.2.3. Fighting the Unregistered Employment

The previous literature on fighting the unregistered employment distinguishes between two broad policy approaches: the dominant deterrence approach that seeks to detect and punish non-compliance; and an emergent approach focused on positively encouraging compliant behavior (Eurofound, 2009). The methods have been named in various ways; i.e., “stick vs carrot”, “chauvinistic” versus “softy”, “command and control” versus “responsive regulation”, etc.

Firstly, “deterrence” approach is a more common and straightforward way to eliminate the informality. As its name implies, deterrence approach includes improving the inspection mechanism and increasing the punishments. This method, in which individuals are viewed as rational actors, constitutes a ‘negative reinforcement’ approach that seeks to elicit a change in behavior using punitive measures for those engaged in non-compliant or ‘bad’ behavior, so that they will change their actions (Eurofound, 2009).

Secondly, “enabling compliance” is a long term approach depending mostly on conscious individuals which are viewed as social actors. It focuses on preventing and curing the incompliance by methods including simplifying the conditions for declaring, applying tax incentives, offering amnesties, providing guidance and support services, etc. Also this method aims to foster the commitment of individuals to social security by educating people while promoting the benefits of registered employment (Eurofound, 2009).

4.2.4. The Unregistered Employment in Turkey

As the World Bank (2010) states, informality is a part of everyday life and nested with formality in Turkey. There is a belief that everybody is engaged to informality in a sort of way. Similarly, another common belief is that public services are taken in low quality and therefore the cost of the service received is already paid.

The salient inflation phenomenon since 1970s (11.9% (CPI) in 2017) and high unemployment rate (10.9% in 2017) make Turkish economy exposed to unregistered employment. On the other hand, sectors with intensive unregistered employment are strong in Turkish economy. Namely, 19.4% of those who were employed in 2017 were employed in agriculture, while it was 19.1%, 7.4% and 54.1% for industry, construction and service sectors respectively according to Turkish Statistical Institute’s (Turkstat) Household Labour Force Survey (HLFS). Furthermore, Small and Medium-Sized Enterprises (SMEs), which are prone to employ without declaring to Social Security Institution, constituted 99.8% of total number of enterprises and 73.5% of employment in 2016 according to Turkstat’s SME Statistics. In addition to these fundamental indicators, increasing population and income inequality give hints regarding the presence of unregistered employment in Turkey.

Though the level of the unregistered employment is decreasing for a long period, it is still very high in Turkey. According to the Turkstat’s HLFS, the ratio of persons working without any social security relating to the main job is realized as 33.4% in

2018. In the non-agricultural sector, on the other hand, the rate of the unregistered employment is realized as 22.3%.

Table 4.1. *Unregistered Employment Numbers and Rates in Turkey (1988 - 2018)*

Year	Number of Employed (,000)	Number of Unregistered Employed (,000)	Unregistered Employment Rate	Unregistered Employment Rate in Agriculture	Non-agricultural Unregistered Employment Rate
1988	17,755	10,320	%58.1	%93.5	%27.4
1993	18,679	8,757	%46.9	%78.2	%25.4
1998	22,334	11,306	%50.6	%87.9	%23.6
2003	21,147	10,943	%51.7	%91.2	%31.5
2008	21,194	9,220	%43.5	%87.8	%29.8
2013	25,524	9,379	%36.7	%83.3	%22.4
2014	25,933	9,069	%35.0	%82.3	%22.3
2015	26,621	8,937	%33.6	%81.2	%21.2
2016	27,205	9,111	%33.5	%82.1	%21.7
2017	28,189	9,575	%34.0	%83.3	%22.1
2018	28,738	9,604	%33.4	%82.7	%22.3

Source: TurkStat, Household Labour Force Survey, 1988 - 2018

Simply by focusing on Table 4.1, one can come to the inference that there is not a dramatic change in the number of people working unregistered and the decrease in the unregistered employment rate arise from the increase in the number of people employed. On the other hand, the World Bank (2010) associates the decrease in the unregistered work with the disengagement from agriculture.

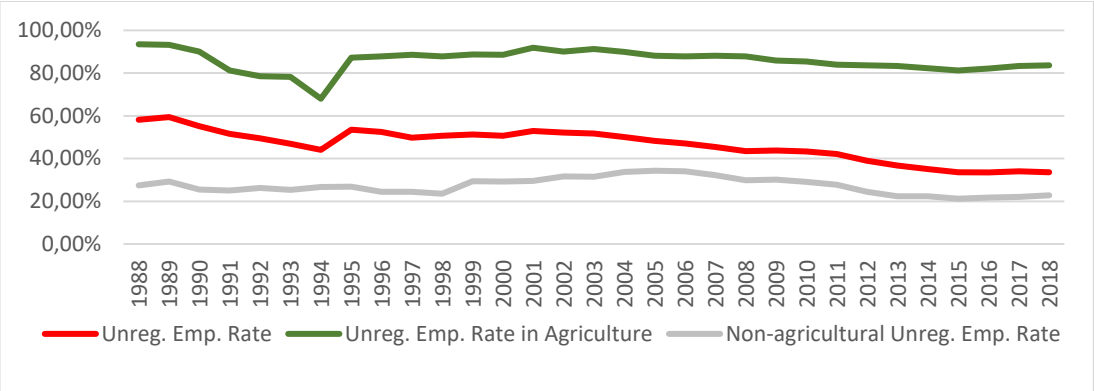

Figure 4.1. The Trend of Unregistered Employment in Turkey (Turkstat, HLFS, 1988 - 2018

4.2.5. Characteristics of Unregistered Employment in Turkey

Focusing on the unregistered employment in Turkey in terms of micro-determinants, some specificities of employees and the workplaces are observed to be critical. This provides important extend of information regarding the characteristics of the unregistered employment in Turkey, and gives the opportunity to compare the phenomenon with the rest of the World in terms of various aspects. In this respect, current situation of Turkey in terms of its micro-determinants of unregistered employment is given briefly with considering the global situation as follows:

4.2.5.1. Gender

As can be seen in Figure 4.2, the differences in unregistered employment rate between women and men change dramatically throughout the World. Even though the global unregistered employment rate is lower for women (58.1%) than men (63%), this rate is higher for women in low and lower-middle income countries. Namely, the unemployment rates are higher for women than men in the most of the sub-Saharan, Southern Asian and Latin America countries. It should also be noted that the women's engagement in the unregistered employment is usually observed in the worst forms (ILO, 2018).

A glance at Turkey case makes it clear that the principal breadwinner role attributed to men is an important factor creating the gender disparity in the local labour market. Even the traditional social security system is established based on a patriarchal family model in which women are the passive beneficiaries. However recent reforms started to create a more gender neutral system while providing the individualization of benefits (Kılıç, 2008).

Turkstat's HLFS puts forth that this way of understanding causes the unregistered employment rate of women (45%) to be realized significantly higher than the rate of men (29%) in Turkey in 2017. Especially in agriculture sector, dramatically large parts of women work unregistered (94%).

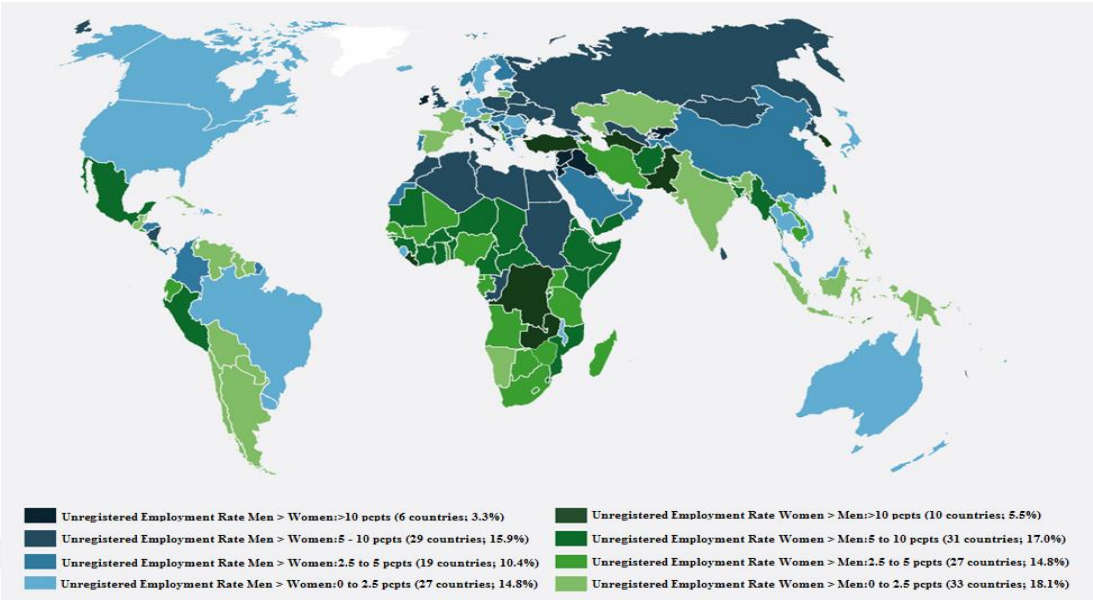

Figure 4.2. The Difference of Unregistered Employment Rates by Genders (ILO, 2018)

4.2.5.2. Age

Unregistered employment is more common for older and younger age groups. According to Figure 4.3, only 22.9% of the aged between 15-24 and 21.2% of the people aged over 65 work registered across the World. Even though, the pattern is similar for emerging and developing countries with developed countries, the rate difference for the same groups are obviously in favor of developed countries.

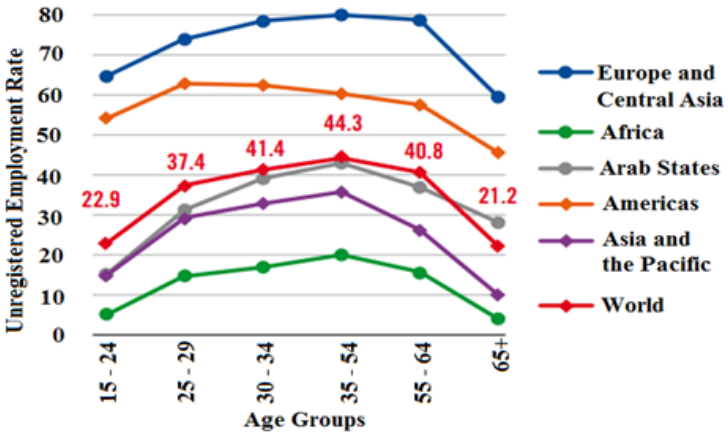

Figure 4.3. Global Registered Employment Rates by Different Age Groups (ILO, 2018)

In line with the global instance, the histogram in Figure 4.4 reflecting the unregistered employment rate by age groups is bi-modal (U-shaped) in Turkey where the two modes are at the lowest and highest age groups.

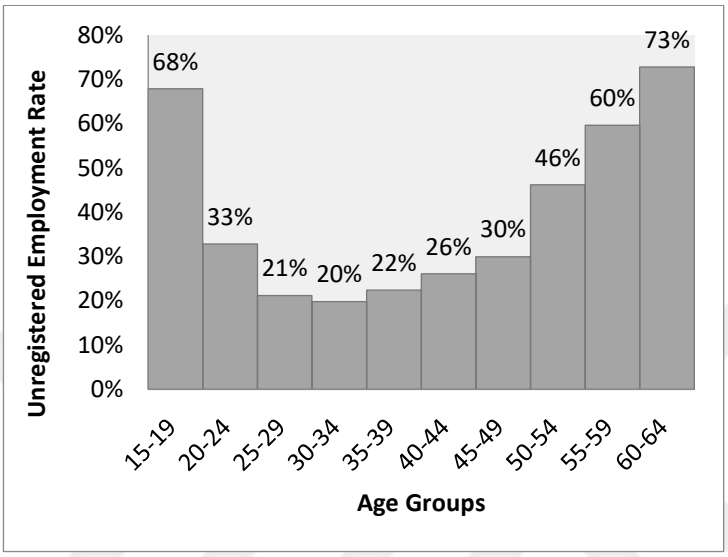

Figure 4.4. Histogram of Unregistered Employment Rates by Different Age Groups (Turkstat, HLFS, 2017)

It can be inferred that the oldest age group benefiting the pension system prefer to turn to labour market with high rates of unregistered employment and thus make a better and easier living after retirement.

In the earliest age group at the other end of the graph, unregistered employment is quite high. Moreover, the rate of informality in this group is well above the age group closer to it. Factors such as the convenience of being in the social security protection dependent to their parents and the tendency for flexible working methods are effective in this situation.

4.2.5.3. Education Status

As ILO (2018) indicates, educational attainment is important in terms of the unregistered employment. Globally, level of education and unregistered employment are inversely related. The likelihood of a person with lower education to work undeclared is higher. As can be seen in Figure 4.5, this relationship is applicable not only for developed countries but also for emerging and developing countries.

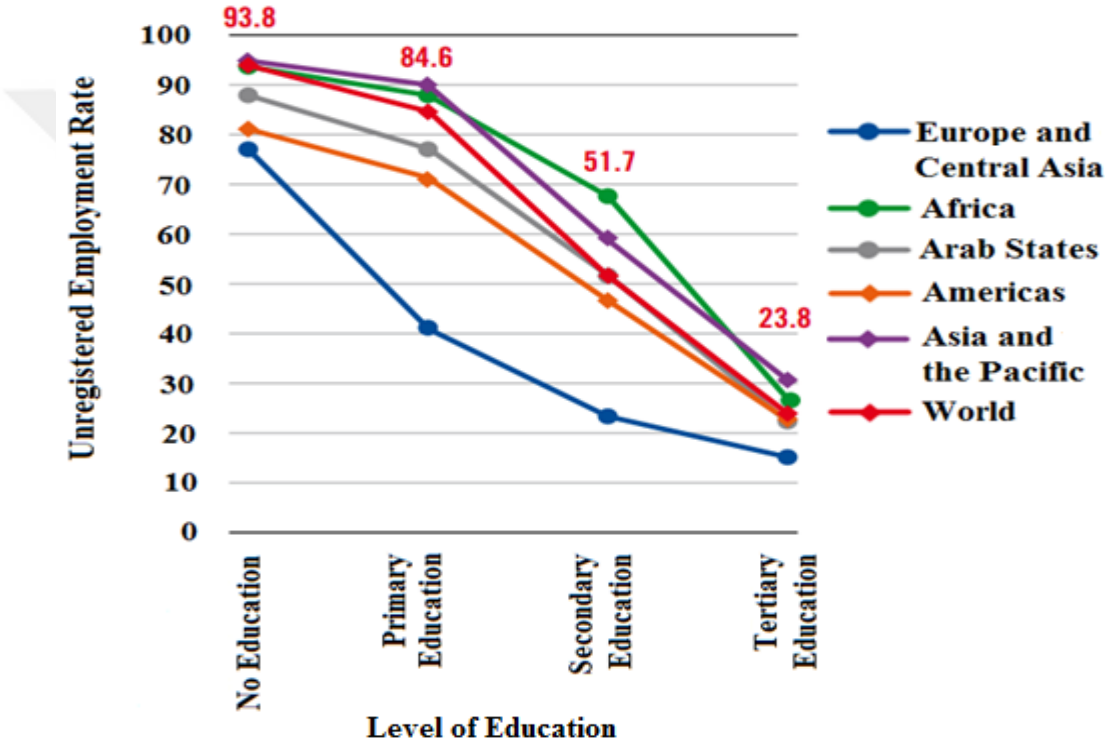

Figure 4.5 Global Registered Employment Rates by Different Levels of Education (ILO, 2018)

In line with ILO's research (2018), this pattern can also be observed when the unregistered employment rate is examined according to the education level in Turkey. According to Figure 4.6, as the level of education increases, the rate of informal employment decreases. This can be associated to educated groups' higher social security awareness and the characteristics of their work.

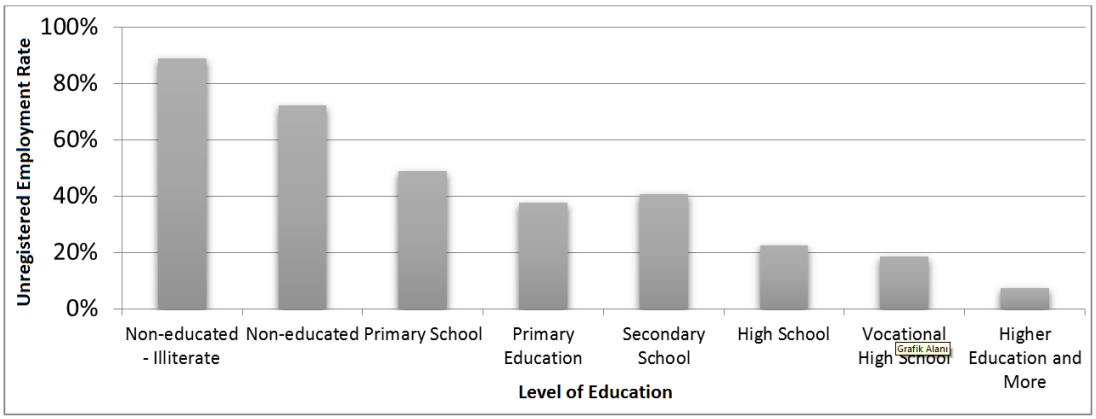

Figure 4.6. Unregistered Employment Rates by Different Levels of Education in Turkey as of 2017 (Turkstat, HLFS, 2017)

4.2.5.4. Marital Status

Considering the property of social security system to provide benefits to persons over their spouses, it can be inferred that marital status is a factor for one's tendency to unregistered employment. Furthermore, the responsibility of having a family can change the attitude of the people towards the informality. Accordingly, unregistered employment rates of people with different marital statuses are calculated from the microdata of Turkstat's HLFS in 2017 within the scope of the thesis to give hints in this context. Based on the calculations, unregistered employment rates for the ones never married, married and divorced are 31.6%, 34.3% and 28.9% respectively. On the other hand, this rate boosts to 75.7% for widows.

4.2.5.5. Sector

Unregistered employment is substantially more prevalent in agriculture sector in the World. 93.6% of the total employment is unregistered while it is 57.2% and 47.2% for the industry and service sectors; respectively (ILO, 2018). Considering employment by sectors in Turkey, agriculture is the main source of unregistered employment (83.3% in 2017) as well. On the other hand, the share of unregistered employment in

manufacturing (20%) and service (21%) sectors are very close to each other, while more than one in every three employees (36%) in the construction sector is working as unregistered.

4.2.5.6. Number of Employees Employed in the Workplace

Depending on the differing level of inspectability, professionalism and institutionalism, unemployment rate is more common in small and middle enterprises (Güloğlu, 2005; DPT, 2001, Sarılı, 2004). Accordingly, unregistered employment rates for workplaces with different number of employees are calculated from the microdata of Turkstat's HLFS in 2017 within the scope of the thesis. As per the calculations, more than half of the employees (56%) in workplaces with 10 or less employees work unregistered. As can be seen in the Figure 4.7, the unregistered employment and the number of persons employed in a workplace are inversely related with each other.

Figure 4.7. Unregistered Employment Rate by Numbers of Employees in the Work Place (Turkstat, HLFS, 2017)

4.2.5.7. Employment Status

Employment status is another important micro-factor determinant on the unregistered employment. Globally, 39.7% of employees, 50.7% of employers and 86.1% of own account workers work without declaring (ILO, 2018). As shown in Figure 4.8, there is a dramatic difference between employment statuses in terms of the unregistered employment in Turkey. While almost all unpaid family workers work unregistered, this is merely rare (18.6% in 2017) in regular employees.

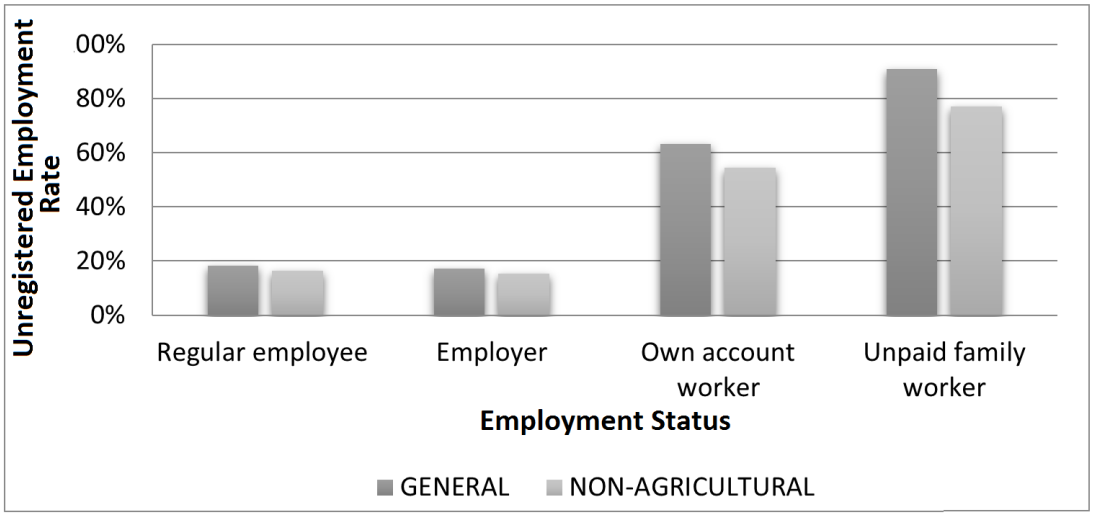

Figure 4.8. Unregistered Employment Rates by Different Employment Status as of 2017 (Turkstat, HLFS, 2017)

4.2.5.8. Type of Employment

Part time employers are more inclined to unregistered employment. 78.7% of those who are working less than 20 hours, 75.1% of those who are working less than 35 hours and 56.5% of those who are working more than 35 hours in a week work unregistered in the World (ILO, 2018). Similarly, unregistered employment rates of 28.4% and 81.6% among full time and part time employees; respectively, in Turkey, indicate a substantial difference with respect to deprivation of social security benefits.

4.2.5.9. Region

Unregistered employment in Turkey, poses regional differences (Güloğlu, 2005; Levent et al. 2004). Figure 4.9, showing the unregistered employment rates for NUTS-II level regions by means of colour representations, is prepared based on the calculations according to microdata of TurkStat's HLFS in 2017. Figure 4.9 shows that unregistered employment increases gradually eastward.

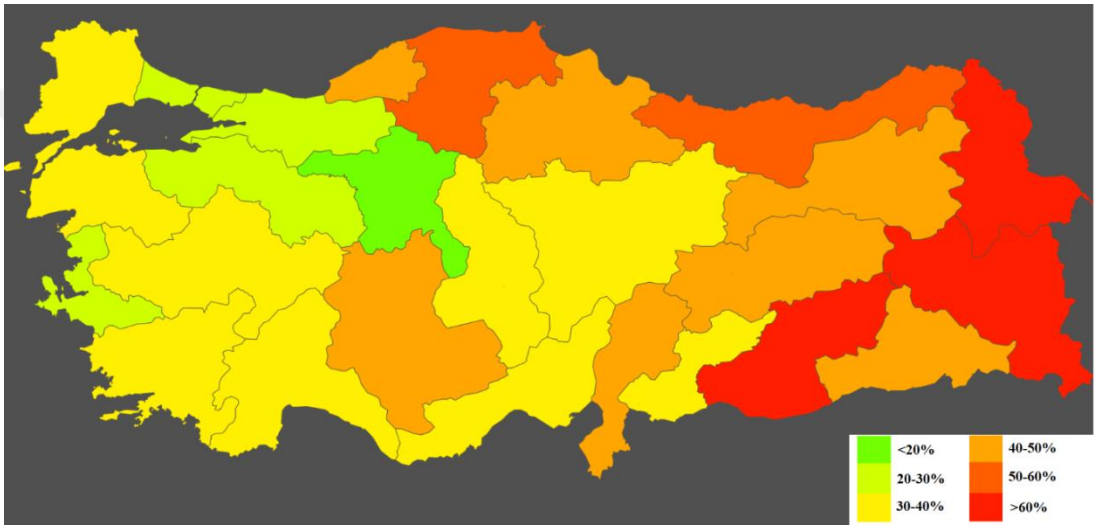

Figure 4.9. Unemployment Rates by NUTS-II Regions (Turkstat, HLFS, 2017)

In depth analysis shows that Ankara and İstanbul are the NUTS-II Regions with the lowest unregistered employment rates; 18.8% and 20.7%, respectively. These two NUTS-II level Regions, consisting of two big cities, are followed by TR41 Region comprised of Bursa, Eskişehir and Bilecik with a ratio of 22.7%. On the contrary, highest unregistered employment rates belong to eastern regions including TRA2, TRB2 and TRC2 with 67.5%, 62.8% and 62%, respectively.

4.2.6. Fighting the Unregistered Employment in Turkey

There is not a single authority to tackle unregistered employment in Turkey. However, Ministry of Family, Labour and Social Services and Social Security Institution (SSI) are the most important institutions in this field. Close cooperation of wide range of state institutions and even private institutions is needed in order to struggle against the phenomenon which has various aspects in many fields.

To fight unregistered employment, Turkey is waiving to implement entirely deterrence approach while engaging different forms of enabling (Eurofound, 2013). However most of the measures taken in this field are still conducted in the conventional forms.

Considering the non-agricultural employment rate's decreasing from 32.3% to 22.1% from 2007 to 2017, it can be inferred at the first glance that the measures taken in the field of unregistered employment are working. Among the major measures taken between 2007 and 2017, the activation of SSI can be mentioned. In 2007, the parties of the Turkish Pension System including Pension Fund of Civil Servants, SSI for employees in private sector and Bağ-Kur for the farmers and self-employed gathered under a single roof of SSI. Then, the Social Security and Universal Health Insurance Law No. 5510 entered into force in 2008, including provisions on the fight against the unregistered employment. For instance, according to the 6th paragraph of the 8th article of the Social Security and Universal Health Insurance Law, the commercial banks and public administrations are obliged to cooperate with SSI, collect information about the registry of the citizens and clients to SSI due to their work and report to SSI in case of observation of the unregistered employment.

The inspection mechanism is enforced with a risk-based perspective and the cooperation between relevant institutions is strengthened while the administrative fine due to the unregistered employment is increased to be more deterrent. Also, Social Security Auditors were employed to serve in the provincial directorates of the Social Security Institution in order to increase the effectiveness in provinces. Furthermore, a

hotline (ALO 170) which was activated in 2008 presents the opportunity to report unregistered employment to the relevant government institutions.

The declaration to the SSI is made obligatory for the professional organizations and tax offices regarding the own account workers who registered to the professional organizations and tax offices due to their activities.

“Regulation on Payments of Wages, Premiums and Every Kind of Remuneration through Banks” entered into force in 2008 and obligated the employers of the enterprises with more than 10 employees to make the payments to the employees by means of banks.

The conditions to benefit from old age insurance are revised and the minimum working days requirement is substantially increased according to the 28th Article. Accordingly, the previous early retirement eligibility is eliminated. This is expected to reflect to the unregistered employment because early retirees were prone to work without declaring to SSI since they have already earned social security rights from the work they retired from.

In line with the “enabling compliance approach” in the field of unregistered employment, various projects, campaigns, events, seminars, meetings, workshops etc. were implemented to raise the awareness of the public. Also, trainings regarding the social security were given to students, soldiers, social assistance beneficiaries etc. in cooperation with various government institutions including the Ministry of National Education, Ministry of National Defence, Ministry of Health

CHAPTER 5

ANALYSIS OF THE UNREGISTERED EMPLOYMENT IN TURKEY

This chapter analyzes the micro-determinants of unregistered employment in Turkey by means of logistic regression analysis after reviewing the previous similar studies in the literature. Detailed comparisons among groups including genders, regions and years are conducted through predicted probabilities as well as average marginal effects and the results are interpreted.

5.1. Aims of Modelling

There are three main aims of the thesis including to reveal the differences in terms of the micro-factors determinant for the occurrence of unregistered employment in Turkey between regions, genders and years. Rather than the macro indicators effective on unregistered employment rate, this study focuses on the individuals and purposes to reach the details behind individuals' preference or obligation to work unregistered. The complex background of unregistered employment will be shed light by means of determinants carefully selected based on expertise on the field. Also, it should be noted that the data is prepared by creating groups from the beginning considering the heterogeneity problem, rather than building one general model by including year, gender and region as independent variables.

The results of the models can be utilized to assign risks to persons in terms of unregistered employment or revised based on workplaces and can be applied prior to relevant inspection of the government institutions responsible in the labour market field in order to provide cost and time efficiency. Considering the disposition of these institutions to conduct risk based inspection, this study can be assessed to be instructive in practice.

Models within the context of this thesis provide a perspective on how the micro-factors determinant on unregistered employment changes according to gender. In this respect, female and male gender groups are compared; the differences are tested in terms of significance and interpreted accordingly. Noting that the gender mainstreaming is a crucial topic for researchers in labour markets, clearly disclosing the disadvantages, mostly suffered by females, based on applications of statistical methods to real data will contribute substantially to the studies in this field.

Similar to the gender perspective, the regional disparities are focused in terms of unregistered employment. For this aim, NUTS-I level regions are accumulated under two classes including East and West in this thesis. The differences among regional groups are tested and interpreted accordingly.

Lastly, models for different years are compared to reach meaningful differences over time. For this purpose, models depending on the data for 2007 and 2017 are built. The reason behind the selection of 2007 is the start of legislative and institutional transformation in the fight against informal employment this year. On the other hand, the most up-to-date data as of the date of drafting this thesis belongs to 2017. Considering that the measures taken in the field of unregistered employment do not produce results in a very short period, it is considered reasonable to evaluate the measures after a period of 10 years. According to the results, the changes over time are associated with the measures taken in the field of unregistered employment where possible.

5.2. Review of the Previous Similar Studies in Literature

There are various studies focusing on the salient unregistered employment phenomenon in Turkey. Majority of these studies address its effects and reasons, measures its extent as well as propose political and technical solutions to the problem. However, rather than the macro perspective, micro determinants of the unregistered employment in Turkey is a research area with limited number of studies.

It should be noted in advance that; these studies approach the concept of unregistered employment in different ways. This arises from the different definitions of the unregistered employment because informality can be defined based on enterprises, job type, production, employment (legality) or registry (social security).

As an empirical study based on micro determinants in terms of informality, Aydın et al (2010) conducted multinomial logit modeling in order to estimate the sectoral allocation of individuals, while investigating the dynamics of labour market segmentation in Turkey. For the dependent variable, it was assumed that there were 5 mutually exclusive alternatives that the individuals face including not working, working in the formal sector, working in the informal sector, self-employed/employer and unpaid family worker. Relying on the results, it was seen that increasing level of education had increasing impact on the probabilities of being observed in all the other alternatives than not working, for both male and female. Furthermore, it was observed that being married increased the probability of males working while it had decreasing effect for females. In this study, informality was referred as all workers not registered to social security institutions and the employees in workplaces with 10 or lower employees.

Doğrul (2012) also built multinomial logit model by means of data from Turkstat's Household Budget Survey of 2006 to investigate the informality specific to urban areas in Turkey. It was concluded that the gender, marital status, breadwinner role and education are of crucial importance in selection of public, private and informal sectors.

Kan (2012) built probit models by using the Survey on Income and Living Conditions of Turkstat to analyze unregistered employment while comparing the different definitions of informality in her study which is a collection of three essays. Similarly, while focusing on employee transitions between distinct labour market states including formal and informal salaried, formal and informal self-employed, unemployed and inactive; Tansel & Kan (2012) conducted multinomial logit regressions in order to analyze the effects of individual profiles covering gender age,

education level, work experience, sector, firm size, number of households, having child and rural/urban. The results of the study revealed that the transition from informality to formality is much higher than the reverse transition in line with conventional theory. Also the study revealed several relationships between the likelihood of transitions and the individual characteristics including gender, education, age, household size and sector of the economic activity.

Additionally, Fidan & Genç (2013) focused on investigating the factors affecting unregistered employment in Turkey. In this study, it is concluded that the number of employees in the working place, core activity, work status and age are found to be more effective factors in explaining exposure risk to unregistered employment. Başlevent & Acar (2015) examined the recent trends in unregistered employment in Turkey by conducting probit regression by imposing a gender based perspective. The econometric analysis in this study yielded results which are in line with the theory. It is observed that women are more likely to work unregistered even after several determinants are controlled for.

Görmüş (2017) analyzed the effects of different individual and workplace based socioeconomic determinants to youth unregistered employment. In this study, micro and small sized establishments, flexible working arrangements, manufacturing sector, lower education and lower ranked occupations are identified as the factors leading the young people in Turkey to unregistered employment.

Furthermore, Levent et al. (2004) divided the individuals employed into segments as working in the formal – informal market or registered-unregistered. In order to investigate the effect of the segmented structure of the labor market on incomes, different models are built by taking income as dependent variable while micro factors belonging to individuals are considered as independent variables.

Bulutay & Taştı (2004) conducted time series regression analysis for distinct definitions of informal employment to reveal their relationship with various macro-level independent variables. Afterwards, micro-level analysis based on individual

based characteristics of the informal employment was conducted without modelling and the effect of migration from rural to urban was highlighted in the conclusion.

In addition to studies focusing on individual or workplace based micro-factors for unregistered employment and informal sector in Turkish labour market, there are similar studies reviewing the other countries in the World. For instance, Williams & Kayaoğlu (2017) conducted a logistic regression analysis at the European Union level in order to investigate the factors effecting the presence or absence of a written employment contract in Europe. The study concluded the unregistered employment in European Union as not associated with socio-demographic or socio-economic characteristics. Rather it was associated with firms' size, institutional and spatial factors.

As a similar but more specific study, Williams & Horodnic (2018) conducted a logistic regression analysis for the presence of employment contract specifically in service sector in Europe (including 28 EU countries, 5 EU candidate countries, Norway and Switzerland). The analysis argued significant associations for various individual and firm based factors including gender, age, education, migration and business size.

Lehmann (2015) conducted a similar study on the micro factors in terms of informal sector in Russia and Ukraine. Similarly, Bracha & Burke (2014) focused on United States while Gasparini &Tornarolli (2009) study on Latin America and Caribbean. Furthermore, Angel & Tanabe (2012) reveal the characteristics of informal employment in the Middle East and North Africa. Also, Radchenko (2014), Sahoo &Neog (2017) and Windebank & Horodnic (2017) focused on Egypt, India and France; respectively.

5.3. Household Labour Force Survey

The models within the scope of this thesis are built by means of the micro data provided by Houshold Labour Force Survey of Turkstat. The main purpose of

Turkstat's Survey can be briefly expressed as revealing the properties and the structure of the Turkish labour force.

Two-stage stratified cluster sampling method is applied in Houshold Labour Force Survey in which household is selected as the statistical unit. Based on address, a rotation pattern is formed to ensure a 50% of overlap between two consecutive periods and in the same periods of the two consecutive years. Yearly sample size of the survey is 176,000 households as of 2014.

The sample selection in this Survey covers all settlements in Turkey. Also, all private households excluding the residents of schools, dormitories, kindergartens, rest homes for elderly persons, special hospitals, military barracks and recreation quarters for officers are covered in labour force surveys. Demographic information (age, sex, educational status, relationship to household head) is asked to all members of the household. But, questions on labour force status are asked for persons 15 years old and over. All information was collected by interviewers on a face-to-face basis.

5.4. Information about the Variables

There are 9 independent variables including gender, age, education status, marital status, sector, number of employees employed in the workplace, employment status, type of employment and region considered as the micro-factors determinant on the choice or obligation of individuals in the labour market to work without declaring Social Security Institution. This consideration is made depending on the relevant literature and the current status of Turkey in terms of unregistered employment. In this respect, characteristics, current status and the trends of unregistered employment in Turkey in terms of selected independent variables are presented in section 4.2.5 of this thesis.

The dependent variable of the models is the social security registry of the employed individual due to his/her main work during the reference week. The options of the

reply include only “yes” and “no”. Accordingly, in line with the relevant question in Turkstat’s Household Labour Force Survey, the quantitative methods in this thesis will treat unregistered employment as not declaring the labour contract to Social Security Institution, while ignoring partially declarations in terms of income and working days, in other words “underreporting”, within the scope of this thesis.

It should be noted that the models within the scope of the thesis are built depending on the non-agricultural data which means that the samples are employed individuals other than the ones working in agriculture sector because more than 4 of every 5 employments in agriculture is unregistered.

The first independent variable is region. Data is categorical in this case and 12 regions are determined based on the 1st level of Nomenclature of Territorial Units for Statistics (NUTS-I) which refers to subdivision standards developed by European Union. The categories of the independent variable include TR1, TR2, TR3, TR4, TR5, TR6, TR7, TR8, TR9, TRA, TRB and TRC. The relevant details are presented in Table 5.1.

The second independent variable is gender and its categories are male and female. During the comparisons of regression models for gender groups, gender is not used as an independent variable accordingly. Rather, one model for each gender is built.

The third independent variable is age. It is the only continuous variable among all independent variables. Since the samples are employed individuals, the ages of the samples are higher or equal to 15 considering the legislation in force regarding the working age. Furthermore, considering the “V shaped” bimodal structure of the unregistered employment among age groups which indicates the higher rate of unregistered employment for younger and older age groups, the square of the age is also included to the model.

Table 5.1. *Categories of the Region Independent Variable and the Provinces under each category*

NUTS-I Level Regions	NUTS-II Level Subregions	NUTS-II Level Provinces
TR1 - Istanbul Reg.	İstanbul Subreg.	İstanbul
TR2 -West Marmara Reg.	Tekirdağ Subreg.	Tekirdağ, Edirne, Kırklareli
	Balıkesir Subreg.	Balıkesir, Çanakkale
TR3 - Aegean Reg.	İzmir Subreg.	İzmir
	Aydın Subreg.	Aydın, Denizli, Muğla
	Manisa Subreg.	Manisa, Afyonk., Kütahya, Uşak
TR4 - East Marmara Reg.	Bursa Subreg.	Bursa, Eskişehir, Bilecik
	Kocaeli Subreg.	Kocaeli, Sakarya, Düzce, Bolu, Yalova
TR5 - West Anatolia Reg.	Ankara Subreg.	Ankara
	Konya Subreg.	Konya, Karaman
TR6 - Mediterranean Reg.	Antalya Subreg.	Antalya, Isparta, Burdur
	Adana Subreg.	Adana, Mersin
	Hatay Subreg.	Hatay, Kahramanmaraş, Osmaniye
TR7 - Central Anatolia Reg.	Kırıkkale Subrg.	Kırıkkale, Aksaray, Niğde, Nevşehir, Kırşehir
	Kayseri Subreg.	Kayseri, Sivas, Yozgat
TR8 - West Black Sea Reg.	Zonguldak Subreg.	Zonguldak, Karabük, Bartın
	Kastamonu Subreg.	Kastamonu, Çankırı, Sinop
	Samsun Subreg.	Samsun, Tokat, Çorum, Amasya
TR9 - East Black Sea	Trabzon Subreg.	Trabzon, Ordu, Giresun, Rize, Artvin, Gümüşhane
TRA - Northeast Anatolia Reg.	Erzurum Subreg.	Erzurum, Artvin, Bayburt
	Ağrı Subreg.	Ağrı, Kars, Iğdır, Ardahan
TRB - Central East Anatolia Reg.	Malatya Subreg.	Malatya, Elazığ, Bingöl, Tunceli
	Van Subreg.	Van, Muş, Bitlis, Hakkari
TRC - Southeast Anatolia Reg.	Gaziantep Subreg.	Gaziantep, Adıyaman, Kilis
	Şanlıurfa Subreg.	Şanlıurfa, Diyarbakır
	Mardin Subreg.	Mardin, Batman, Şırnak, Siirt

The fourth independent variable is named as education and refers to the level of education. It has 6 categories including non-educated persons, primary school graduates, secondary school graduates, high school graduates, vocational and technical high school graduates, graduates of higher education or over.

The fifth and the sixth independent variables are marital status and employment status. There are 4 categories of marital status covering never married, married, divorced and widow. On the other hand, employment status has also 4 categories including regular employee, employer, own account worker and unpaid family worker.

The seventh independent variable is sector and refers to the main activity of the workplace where the sample is working. In TurkStat's Household Labour Market Survey, the classification is made according to industrial standard classification system of European Union which is known as Statistical Classification of Economic Activities in the European Community (NACE). Within the scope of this study, NACE codes (from 01 to 99) of the samples are consolidated under 4 categories including agriculture, manufacturing, service and construction. Since the study is based on non-agricultural employment, samples in agriculture sector are excluded accordingly. Namely, sector independent variable is comprised of 3 categories including manufacturing, service and construction.

The eighth independent variable is the number of employees in the workplace where the sample is working. It has 3 categories covering less than 11, between 11 and 49 and more than 49 which can be evaluated as small, medium and large enterprises from another perspective.

Finally, the ninth independent variable is the type of employment with 2 categories covering full time employment and half time employment.

5.5. Econometric Analysis

Econometric analyses are categorized under 3 main topics in this thesis. The first part of the analysis is conducted based on a 10-year time period. In this context, logit models for 2007 and 2017 are built. Afterwards, the results of the model for 2017 are interpreted in this part in order to give the initial insight regarding the micro-level determinants of the unregistered employment in Turkey. Furthermore, group comparisons between years are made. In this respect, the differences in micro-factors of unregistered employment between 2007 and 2017 are revealed and interpreted according to the results. The differences in micro-factors through time is analyzed and associated with the political measures.

The second part is the analysis of the models for the gender groups. Different logit models for women and men are built and interpreted in this part. Afterwards, the differences between men and women in terms of micro-determinants of unregistered employment are revealed.

The third part is the analysis of the models for the regional groups. Similar to gender perspective, different logit models for eastern and western regions are built and the differences are revealed and interpreted.

It should be noted that the interpretation of the models is made by means of predicted probabilities and average marginal effects in order to benefit from their straightforward nature. For the comparison of predicted probabilities, profiles to be compared are determined in line with the research question.

In order to reach more reliable results, inter-associations between the independent variables are checked to detect multicollinearity. In this respect, the correlation coefficient for the only two continuous variables at the beginning including “age” and “duration of the employment” are calculated as higher than 0.5 and “duration of the employment” is eliminated from the model accordingly.

In addition to the conventional correlation coefficient method, Cramer’s V Coefficient, which is derived from Pearson’s Chi-Squared Test is applied for the same

purpose considering the categorical feature of the rest of the independent variables. Regarding the interpretation of V coefficient, values lower than 0.3 indicate weak association, values between 0.3 and 0.6 indicate moderate association and values higher than 0.6 indicate as strong association. According to the calculations, even the highest V coefficient is lower than 0.5 while most of the V values indicate weak associations between the independent variables.

Interactions are not included to the models consciously. Depending on the consideration of making comparisons as the primary focus of the study rather than interpreting a general model, interactions are neglected in order to keep the simplicity during the interpretations of the comparisons.

Average marginal effects and the predicted probabilities for the comparisons are calculated based on the results of logistic regression models. In this respect, the results of the logistic regression models for 2017, 2007, males, females, eastern regions and western regions are given in Appendices from A to F, respectively.

Analysis in the thesis are completed using R version 3.2.3. In this context, the relevant codes are given in the Appendix-G. However, it should be noted that the sample codes for each step, including the construction of the confidence interval and the calculation of Z-Statistic of probability differences during the comparison is given, rather than giving similar codes repeatedly for each step.

5.5.1. Analysis of the Overall Models for 2007 and 2017

Logistic regression models with the same independent variables are built for the years 2007 and 2017 to further understand changes by time in the micro-factors of unregistered employment in Turkey.

In order to classify the results as registered or unregistered, a cut-off point is determined. For this aim, a graph for the accuracy rates for different cut-off points given in Figure 5.1, is created and 0.5 is determined as the cut-off point for both models in this sense.

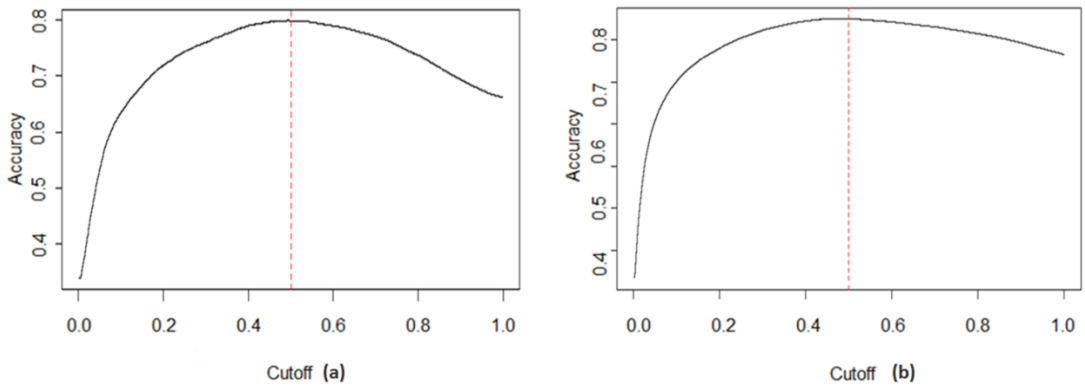

Figure 5.1. Accuracy Rate – Cut-off Point Graphs for the Overall Models of 2007(a) and 2017(b)

Regarding the overall models' goodness of fit, accuracy rates depending on the classification tables are calculated as 79.9% and 85.02% for 2007 and 2017; respectively, which mean the percentage of true classifications over observations. Furthermore, ROC Curves for each model are given in Figure 5.2 for the same purpose while the areas under ROC Curves are also calculated as 0.866 and 0.896 for 2007 and 2017; respectively, which indicates to almost outstanding level of fit.

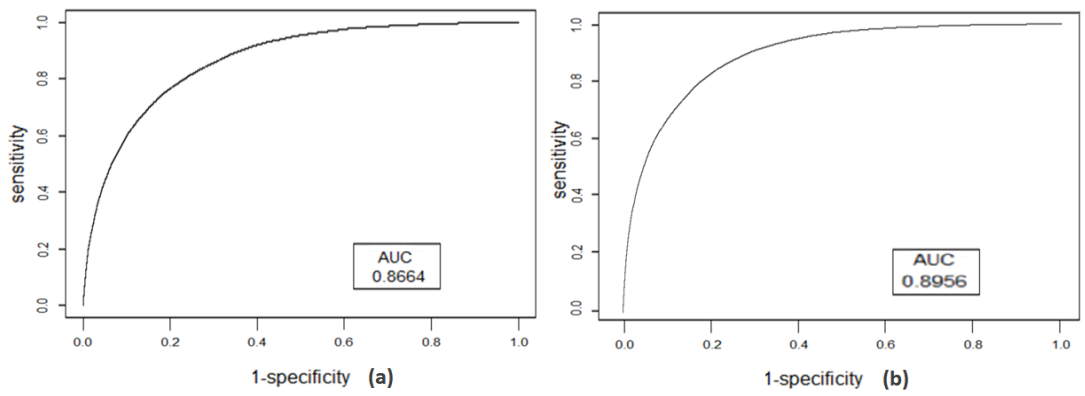

Figure 5.2. ROC Curves for the Overall Models of 2007(a) and 2017(b)

Overall model for 2017 helps to observe the significant micro-factors for the unregistered employment in Turkey. At first glance, the results can be briefly concluded to be in line with the thematic knowledge, hence the expectations. However, model provides the necessary knowledge regarding the strength of the determinants while comparing the categories to each other.

Table 5.2 gives the AME's of the micro-factors of unregistered employment for 2007 and 2017. Standard errors are given in parenthesis, while base categories for the categorical independent variables are female, higher education and more, divorced, employer, construction, between 11 and 49, full time and TR1 Region.

As can be seen from Table 5.2, the average marginal effects of overall model for 2017 put forth that being female is associated with almost 4.9% significant increase in probability of unregistered employment when the other variables are controlled. This can be considered as a clear gender difference in terms of unregistered employment. On the other hand, education is analyzed to be strongly effective in unregistered employment. Controlling for other factors, the probability of a non-educated person to work unregistered is almost 29.5% higher on average than the ones with higher education and more. This difference decreases gradually to 16.5% and 12.3% for graduates of primary school and secondary school; respectively. Besides, the lowest difference with the graduates of higher education and more in terms of probability of unregistered employment belongs to the ones which are the graduates of vocational high schools with 6.3% while it is 7.6% for high school graduates. The results indicate the strong negative relationship between the education level and the probability of unregistered employment. Additionally, attaining vocational high schools can be considered as associated to lower risk of unregistered employment and advantageous compared to regular high schools from this perspective.

Average marginal effects of all marital statuses except widow are statistically significant. The probability of unregistered employment for a married group is the lowest among all. Considering the conventional requirement of a decent job with social security for a marriage, the result is not surprising. Namely, controlling other factors, the risk of unregistered employment for married individuals and never married individuals are 5.5% and 3.7% lower; respectively, than the divorced individuals. On the other hand, the difference is not statistically significant between marital statuses of widow and divorced.

Table 5.2. AME's of the Micro-Factors of Unregistered Employment for 2007 and 2017

	2007	2017
Gender (male)	-0.0588*** (0.0037)	-0.0489*** (0.0024)
Age	-0.0268*** (0.0007)	-0.0284*** (0.0005)
Age-squared	0.0003*** (0.0000)	0.0004*** (0.0000)
Education (non-educated)	0.4042*** (0.0086)	0.2947*** (0.0064)
Education (primary school)	0.2460*** (0.0042)	0.1649*** (0.0030)
Education (secondary school)	0.1906*** (0.0049)	0.1227*** (0.0030)
Education (high school)	0.1054*** (0.0047)	0.0764*** (0.0034)
Education (vocational high school)	0.0929*** (0.0053)	0.0632*** (0.0034)
Marital Status (married)	-0.0804*** (0.0120)	-0.0553*** (0.0057)
Marital Status (never married)	-0.0172 (0.0125)	-0.0370*** (0.0064)
Marital Status (widow)	-0.0446* (0.0186)	-0.0095 (0.0117)
Employment Status (own account worker)	0.1268*** (0.0047)	0.1286*** (0.0041)
Employment Status (regular employee)	0.1742*** (0.0044)	0.0529*** (0.0035)
Employment Status (unpaid family worker)	0.3672*** (0.0085)	0.2368*** (0.0075)
Sector (manufacturing)	-0.1652*** (0.0056)	-0.0303*** (0.0036)
Sector (service)	-0.1948*** (0.0050)	-0.0681*** (0.0030)
Number of Employees (less than 11)	0.2733*** (0.0040)	0.2118*** (0.0029)
Number of Employees (more than 49)	-0.1322*** (0.0033)	-0.0689*** (0.0024)
Type of Employment (part time)	0.2179*** (0.0103)	0.2542*** (0.0046)
Region (TR2)	-0.0416*** (0.0064)	0.0257*** (0.0044)
Region (TR3)	-0.0785*** (0.0042)	-0.0177*** (0.0036)
Region (TR4)	-0.0535*** (0.0047)	-0.0038 (0.0040)
Region (TR5)	-0.0584*** (0.0051)	0.0088* (0.0037)
Region (TR6)	0.0176** (0.0049)	0.0341*** (0.0038)
Region (TR7)	-0.0446*** (0.0067)	-0.0087. (0.0045)
Region (TR8)	-0.0595*** (0.0050)	0.0064 (0.0042)
Region (TR9)	-0.0622*** (0.0071)	-0.0288*** (0.0052)
Region (TRA)	0.0044 (0.0080)	0.0383*** (0.0054)
Region (TRB)	0.0543*** (0.0072)	0.0877*** (0.0050)
Region (TRC)	0.1201*** (0.0061)	0.0989*** (0.0043)

* p<0.05, ** p<0.01, *** p<0.001

Regarding the employment status, employers are the ones with lowest risk of unregistered employment. The risk increases dramatically for unpaid family workers and own account workers. Controlling other factors, the average marginal effects of being unpaid family worker and own account worker are 23.7% and 12.9% when being employer is the base category of employment status. This effect corresponds to 5.3% for a regular employee. It should be taken into account that this study does not include agricultural employment. So, considering the prevalence of unpaid family workers and unregistered employment in agriculture, this risk would increase if agricultural employment was included.

Concerning the sector of the workplace, construction is the riskiest field in terms of unregistered employment. However, the average marginal effect or the difference of probabilities, is not dramatic between the sectors. Namely, service and manufacturing sectors are; respectively, 6.8% and 3% less risky than the construction sector.

Number of employees in the work place as an indicator of institutionalization is another important determinant in the unregistered employment. Controlling for the other variables, the probability of unregistered employment for individuals working in micro and small enterprises with 10 and less employers is 21.2% higher than the ones working in enterprises between 11 and 49 employees. On the other hand, the probability of unregistered employment for individuals working in big enterprises with more than 49 employees is 6.9% lower than the ones working in enterprises between 11 and 49 employees.

The average marginal effects corresponding to each NUTS-I Region of Turkey is critical in order to see the regional differences according to the overall model. Noting that TR1 Region is selected as the base category of the region independent variable, average marginal effects indicate the average differences in unregistered employment probabilities between an individual in TR1 Region and individuals in other NUTS-I Regions while controlling for all other variables.

TRC Region is the NUTS-I Region with the highest probability of unregistered employment. Namely, probability of working undeclared in TRC Region is 9.9% higher than TR1 Region, while it is 8.8% and 3.8% higher than TR1 Region in TRB and TRA regions respectively, controlling for the other variables. On the other hand, the lowest probability of unregistered employment belongs to TR9 Region with 2.9% lower probability of unregistered employment than TR1 Region. Furthermore, it is also 1.8% and 0.3% lower than TR1 Region in TR3 and TR7 Regions; respectively.

5.5.1.1. Revealing the Differences among 2007 and 2017

When 2007 and 2017 models are compared thoroughly, it is obvious that most of the average marginal effects of the determinants changed and most of the differences between the average marginal effects of the categories narrowed by time. For instance, the probability of unregistered employment for men is 5.9% lower on average than women in 2007 and 4.9% lower on average in 2017, controlling for the other variables. Also, it is observed that the effect of education level decreases from 2007 to 2017. Namely, being non-educated increases the probability of unregistered employment by 40.4% on average compared to being a graduate of higher education and more in 2007, while it increases 29.9% on average in 2017. Furthermore, being a graduate of primary school increases the probability of working unregistered by 14.1% and 8.85% on average in 2007 and 2017, controlling for the other variables. Similarly, the differences of average marginal effects for every level of education decreases in 2017 compared to 2007.

From the perspective of marital status, being married decreases the probability of unregistered employment by 6.3% on average compared to being never married in 2007 while it decreases 1.8% on average in 2017. However, the difference in probability of working unregistered between widow and married are close to each other, around 4.4% on average, for 2007 and 2017.

Regarding the employment status, the probability of working unregistered is 4.7% lower on average for own account workers than regular workers in 2007. However, it is 7.6% higher on average in 2017. Furthermore, the difference in probability of unregistered employment of own account worker and employer does not change significantly from 2007 to 2017. Since the declaration to the SSI was made obligatory for the professional organizations and tax offices regarding the own account workers who registered to the professional organizations and tax offices due to their activities, this trend in the category of own account working can be considered as surprising.

Remarkably, the difference of average marginal effects between construction and other main sectors decreased from 2007 to 2017. Working in construction sector increases the probability of unregistered employment by 19.5% on average compared to working in service sector in 2007 while it increases the same probability by 6.8% in 2017. Similarly, the same difference with manufacturing sector in probability is 16.5% on average in 2007 and 3% on average in 2017.

Regarding the effect of the size of the enterprise on unregistered employment, the difference in probability of working unregistered between working in a workplace with less than 11 employees and working in a workplace with employee number between 11 and 49 is 27.33% on average in favor of workplace with employee number between 11 and 49 in 2007. This difference decreases to 21.2% in 2017. The same difference between working in a workplace with less than 11 employees and working in a workplace with more than 49 employees decreases from 40.6% to 28.1% from 2007 to 2017. This decrease can be associated with the “Regulation on Payments of Wages, Premiums and Every Kind of Remuneration through Banks” which entered into force in 2008 and obligated the employers of the enterprises with more than 10 employees to make the payments to the employees by means of banks. It should be noted that the threshold is revised as 5 employees with an amendment in the Regulation in May, 2016.

Regarding the regional disparity over time, average marginal effect differences between NUTS-I Regions narrows remarkably. In 2007, NUTS-I level regions with lowest and highest average marginal effect are TR3 and TRC Regions respectively and the difference of average marginal effect between these two regions is almost 19,9%. On the other hand, TR9 and TRC are the regions with lowest and highest Average Marginal Effects respectively with a difference of 12.8% in 2017.

When the regions ordered in terms Average Marginal Effects, the biggest changes in order happens in TR8, TR5, TR7 and TR1 Regions. When ordered with ascending sort, TR8 and TR5 are ordered 3rd and 4th in 2007, however the same regions are ordered 6th and 7th in 2017 respectively. Conversely, TR7 and TR1 regions are ordered 6th and 8th in 2007 and 3rd 5th in 2017; respectively.

5.5.2. Analysis of Separate Models According to Gender

After fitting an overall model, separate models for female and male individuals are constructed for comparison purposes based on data belong to 2017. Similar to the overall models, cut-off points for the both gender groups are also determined as 0.5 considering the maximum accuracy rate. The cut-off point graphs for both groups are given in Figure 5.3.

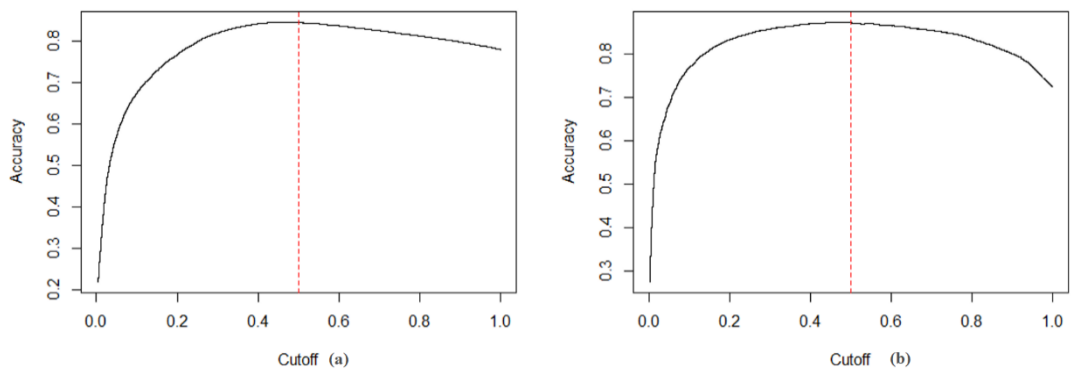

Figure 5.3. Accuracy Rate – Cut-off Point Graph for the Models of Male(a) and Female(b) Groups

As a measure of goodness of fit, accuracy rates for the models built for males and females are 84.6% and 87.2%; respectively, which means that the very large part of the observations classified correctly as registered or unregistered based on the models. On the other hand, the ROC Curves for models of both gender groups are given in Figure 5.4 and the areas under ROC Curves f are calculated as 0.88 and 0.93 for males and females; respectively, which means extensive level of goodness of fit for both models.

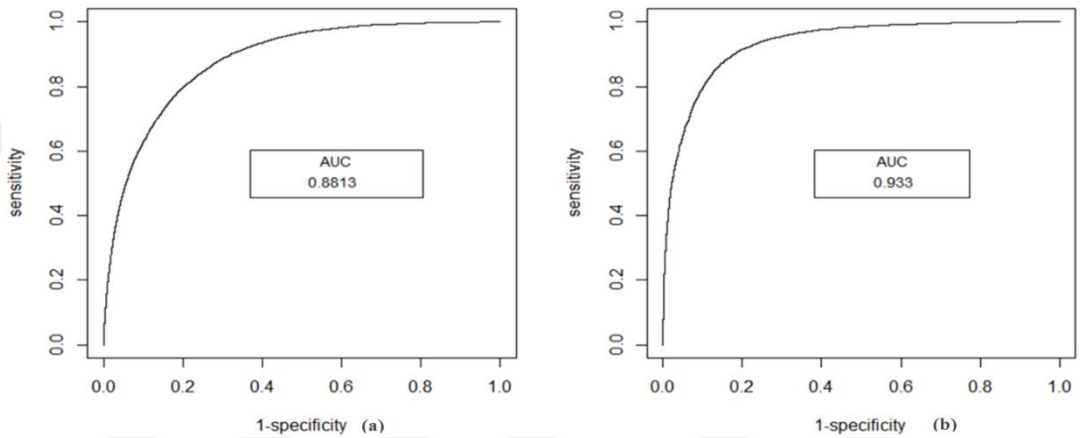

Figure 5.4. ROC Curve for the Models of Male(a) and Female(b) Groups

Table 5.3 gives the AME's of the micro-factors of unregistered employment for males and females. Standard errors are given in parenthesis, while base categories for the categorical independent variables are higher education and more, divorced, employer, construction, between 11 and 49, full time and TR1 Region.

When the Average Marginal Effects in Table 5.3 are analyzed, the effect of education in unregistered employment for women is higher. Namely, for non-educated women, the probability of working unregistered is 34.4% higher than a woman with higher education and more. The difference of probability for the men with same levels of education is 24.6%. Furthermore, being a graduate of high school compared to being a graduate of primary school significantly decreases the probability of unregistered employment by 14.8% and 6.6% on the average for women and men; respectively. To

sum up, the decrease in the probability of working unregistered in return of increasing level of education is usually higher for women.

Regarding the marital status, being married decreases the probability of working unregistered more apparently for men compared to women. The probability of attaining in unregistered employment decreases by 2.6% on the average for married men compared to never married men. However, there is almost no difference in probability for women in this case. Moreover, the average difference in probability for divorced men than married men is 6.9% increase controlling for the other variables while the same difference for women is 2.8% increase. This phenomenon can be associated with the breadwinner role attributed to men and the traditional social security system established based on a patriarchal family model in which women are the passive beneficiaries.

When the average marginal effects of the employment statuses are analyzed for each gender, being a regular employee, which is the largest category of employment status, increases the probability of unregistered employment by almost 5% on the average compared to being an employer for both. Furthermore, being an own account worker and unpaid family worker increases the probability of unregistered employment by 6.9% and 23.1% on the average; respectively, compared to regular employee for men. The same increase corresponds to 13.1% and 10.9% on the average; respectively, for women.

The direction of the average marginal effects regarding the sector of the workplace for men and women are converse at the first glance. However, it should be noted that the reference category is construction in which the women employment is very limited. However, when analyzed more carefully, it can be observed that working in the manufacturing sector increases the unregistered employment probability by 2.8% and 5.2% compared to service sector for men and women, respectively. Accordingly, it may be inferred that women's unregistered employment is more sensitive to the sector in terms of unregistered employment.

Table 5.3. AME's of the Micro-Factors of Unregistered Employment for Males and Females

	Male	Female
Age	-0.0310*** (0.0006)	-0.0207*** (0.0010)
Age-squared	0.0004*** (0.0000)	0.0003*** (0.0000)
Education (non-educated)	0.2457*** (0.0083)	0.3437*** (0.0104)
Education (primary school)	0.1303*** (0.0035)	0.2346*** (0.0060)
Education (secondary school)	0.0974*** (0.0036)	0.1666*** (0.0064)
Education (high school)	0.0640*** (0.0041)	0.0868*** (0.0063)
Education (vocational high school)	0.0510 (0.0040)	0.0734*** (0.0063)
Marital Status (married)	-0.0692*** (0.0091)	-0.0277*** (0.0066)
Marital Status (never married)	-0.0433*** (0.0097)	-0.0289** (0.0082)
Marital Status (widow)	-0.0527* (0.0203)	0.0030 (0.0125)
Employment Status (own account worker)	0.1189*** (0.0044)	0.1855*** (0.0130)
Employment Status (regular employee)	0.0501*** (0.0038)	0.0544*** (0.0115)
Employment Status (unpaid family worker)	0.2819*** (0.0114)	0.1632*** (0.0136)
Sector (manufacturing)	-0.0519*** (0.0040)	0.1013*** (0.0148)
Sector (service)	-0.0797*** (0.0033)	0.0490** (0.0142)
Number of Employees (less than 11)	0.1925*** (0.0033)	0.2413*** (0.0058)
Number of Employees (more than 49)	-0.0683*** (0.0028)	-0.0722*** (0.0051)
Type of Employment (part time)	0.2813*** (0.0069)	0.1889*** (0.0059)
Region (TR2)	0.0231*** (0.0054)	0.0237* (0.0072)
Region (TR3)	-0.0277*** (0.0044)	0.0005 (0.0060)
Region (TR4)	-0.0091. (0.0049)	-0.0022 (0.0067)
Region (TR5)	-0.0025 (0.0045)	0.0304*** (0.0064)
Region (TR6)	0.0238*** (0.0046)	0.0473*** (0.0063)
Region (TR7)	-0.0327*** (0.0052)	0.0493*** (0.0086)
Region (TR8)	-0.0094. (0.0051)	0.0317*** (0.0072)
Region (TR9)	-0.0463*** (0.0061)	0.0130 (0.0098)
Region (TRA)	0.0250** (0.0063)	0.0590 (0.0108)
Region (TRB)	0.0769*** (0.0058)	0.1033*** (0.0100)
Region (TRC)	0.0930*** (0.0051)	0.1096*** (0.0087)

* p<0.05, ** p<0.01, *** p<0.001

Regarding the marginal effect of number of employees in the enterprise, controlling for other factors, the probability of working unregistered is 24.1% higher for women working in micro or small enterprises with less than 11 employees compared to women working in middle enterprises with number of employees between 11 and 49. The corresponding increase in probability is 19.3% for men.

The marginal effect of the type of employment is also different for genders. Controlling for other factors, while the probability of unregistered employment increases 28.1% for men, it increases 18.9% for women.

Interpreting the average marginal effects in terms of NUTS-I Regions is more straightforward by putting in order in terms of probabilities of unregistered employment. Inferring from the orders, while controlling other factors; both men and women are of highest probabilities in TRC, TRB and TRA Regions; respectively. On the other hand, men in TR9, TR7 and TR3 Regions and women in TR4, TR1 and TR3 Regions are of the lowest probability of unregistered employment; respectively. Furthermore, the difference in probability of unregistered employment between the regions associated with highest and lowest probabilities (TRC and TR9 for men and TRC and TR4 for women) are 11.1% and 13.9% for men and women; respectively.

5.5.2.1. Revealing the Differences among Genders

Group comparisons in this study are conducted in terms of predicted probabilities in line with the Long's (2009 and 2018) method. For this aim, the probabilities of unregistered employment for specific profiles with the same micro-factors for each gender are compared. Specific profiles are determined considering the representative categories for each independent variable. Furthermore, the significance of the differences in the predicted probabilities is questioned. The profiles determined are married and residing in İstanbul, working as full-time regular employees in enterprises with less than 11 employees in the service sector. The comparison is made through ages of the genders and conducted for different education levels.

In this context, the difference in probability of unregistered employment is observed to decrease gradually with increasing level of education. After a certain extent, surprisingly, the probability of unregistered employment for men starts to exceed the same probability for women. It can be easily concluded that women are able to escape from informality by increasing level of education.

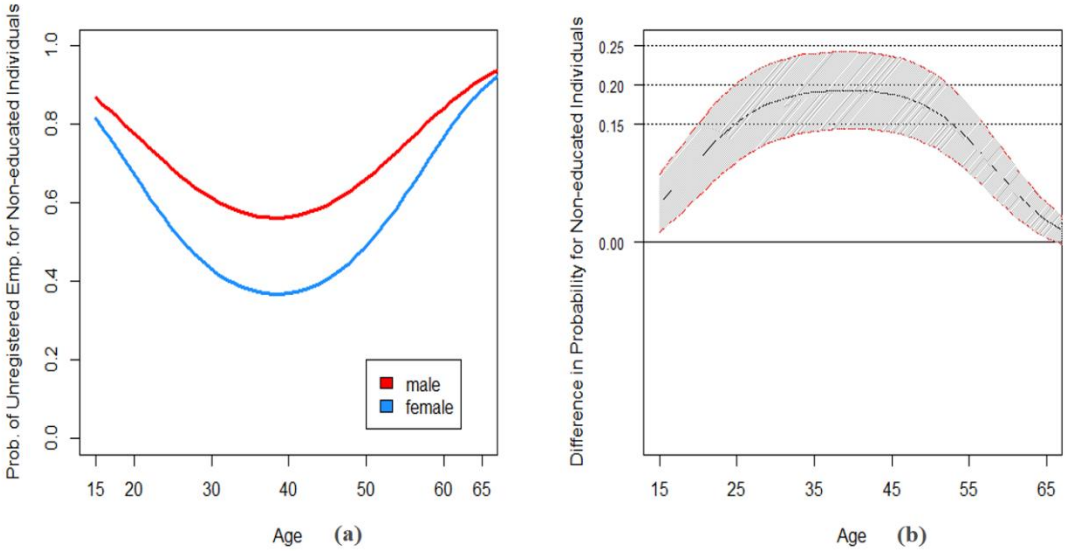

Figure 5.5. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Non-Educated Individuals

As can be seen in Figure 5.5, the difference in probability is highest in favor of men for non-educated groups. The graph of difference is inverted U-shaped and the difference is at its peak, close to 20%, around 39 years old while the difference is significant at a 0.05 significance level, over the whole age period from 15 to 65.

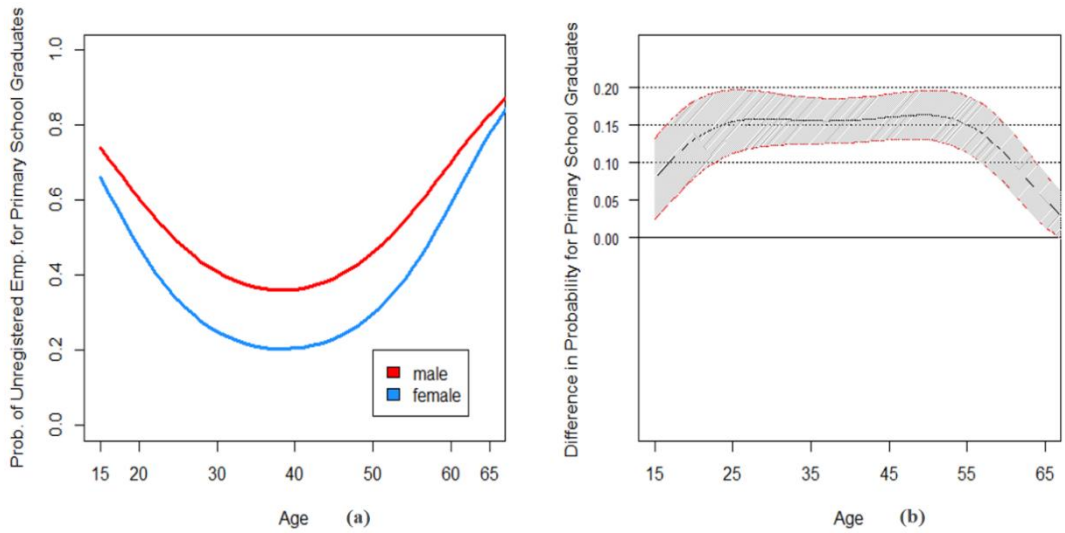

Figure 5.6. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Primary School Graduates

Figure 5.6 indicates similar results observed for the primary school graduates with lower amount of difference in probability while it is still significant from 15 to 65 years old. The difference of unregistered employment for genders remains close to the level of 15% from 20 to 55 years old.

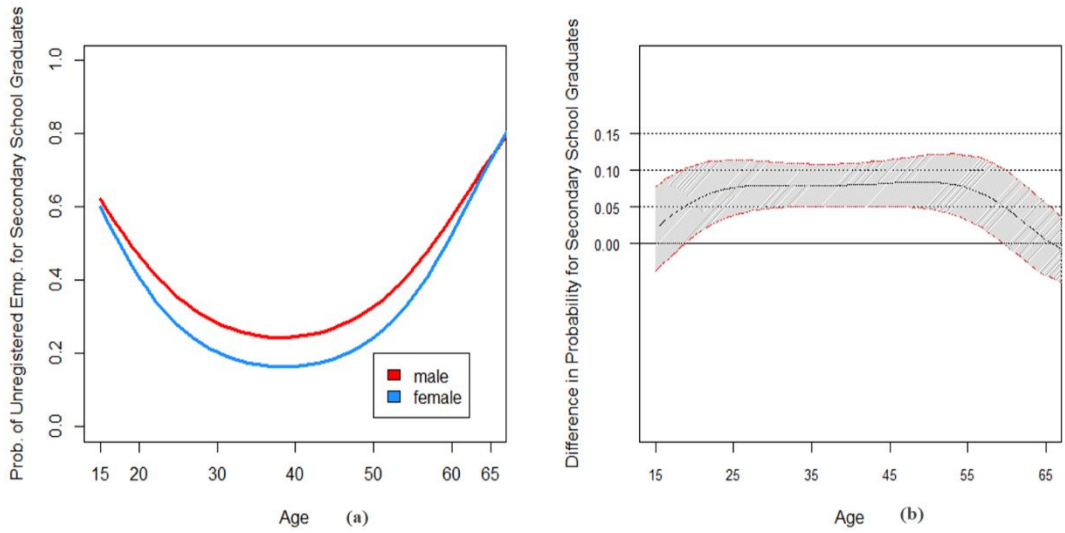

Figure 5.7. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Secondary School Graduates

For the secondary school graduates, the difference in probability of unregistered employment decreases to the level of 7-8% with similar shape to primary school graduates as indicated in Figure 5.7.

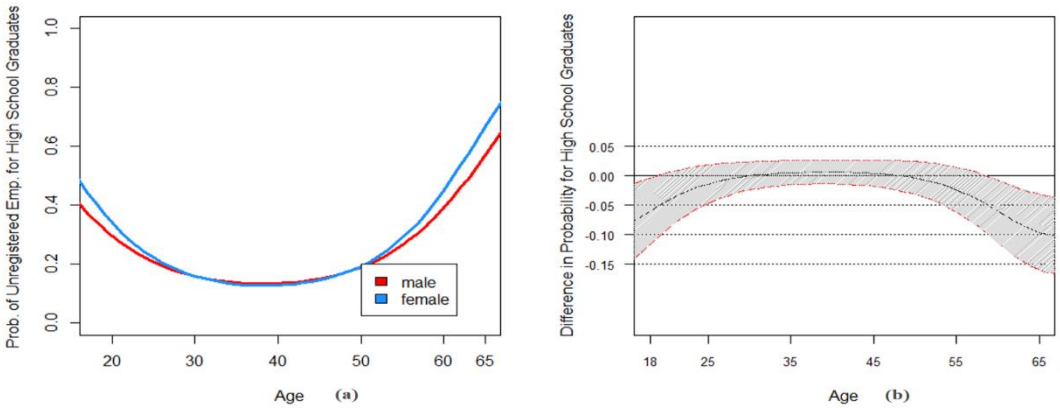

Figure 5.8. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for High School Graduates

As can be seen in Figure 5.8, the direction of the difference in probability changes through the age as a breaking level for the high school graduates. For those younger than 30 and older than 49, the difference in probability turns out to be in favor of women while it is very close to zero for middle ages. From the significance perspective, the difference in probability can be concluded as insignificant almost through whole age period in this education level.

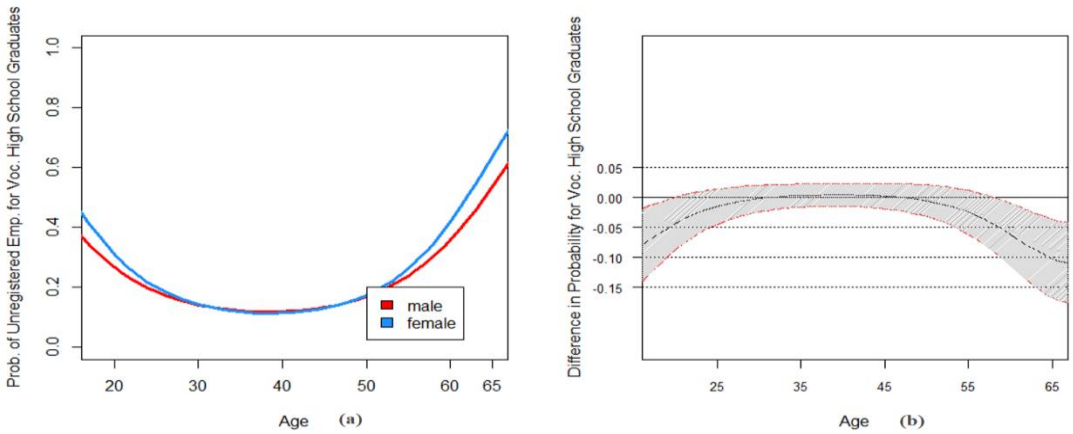

Figure 5.9. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Vocational High School Graduates

Even though it is similar to high school graduates, the advantageous situation of women is more obvious for vocational high school graduates as indicated in Figure 5.9.

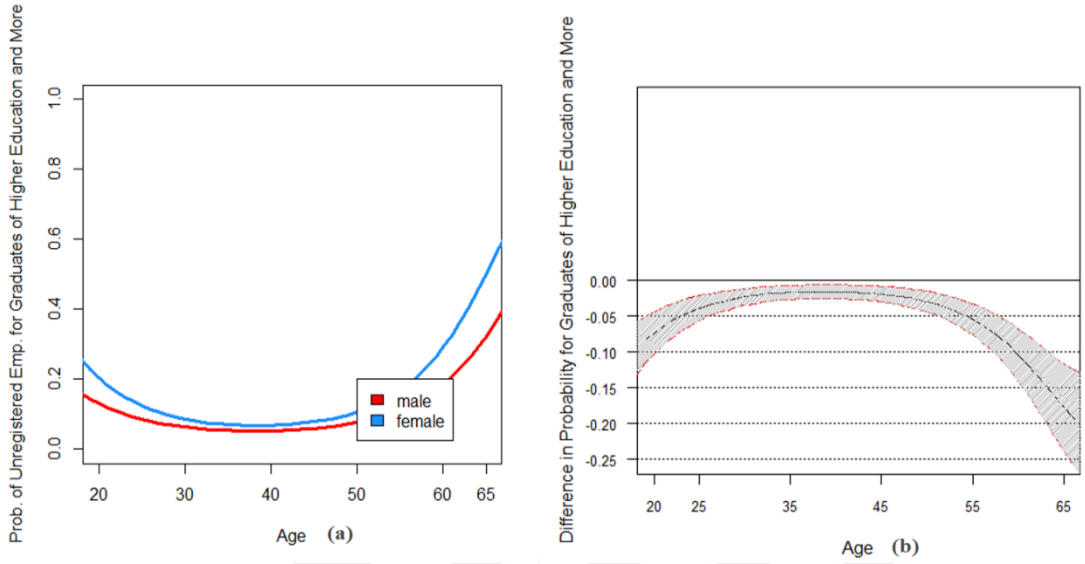

Figure 5.10. Probabilities of Unregistered Employment for Each Gender at Different Levels Through Age (a) and the Differences in Probabilities between Genders with Confidence Intervals (α=0,05) (b) for Graduates of Higher Education and More

As a remarkable point, the probability of unregistered employment of women with higher education and over is significantly lower than that of men, and this is consistent over the whole age period from 15 to 65 as indicated in Figure 5.10. During the middle ages from 25 to 45, the difference remains between 1.8% and 4%.

5.5.3. Analysis of Separate Models According to Region

In order to obtain a better understanding of regional disparity in terms of unregistered employment in Turkey, different models are built for different regions based on data belong to 2017. For this aim, NUTS-I level regions in the dataset are accumulated under the topics of “East” and “West”. In this context, TR1, TR2, TR3, TR4, TR5, TR6, TR7 and TR8 regions are accounted as West; while TR9, TRA, TRB and TRC

regions are accounted as East. Therefore, the difference of the micro-factors for unregistered employment between the eastern and western regions is focused.

Since the interpretations are made upon the predicted probabilities and the average marginal effects, the cut-off point for the classification of the results as registered and unregistered is important. Similar to the overall model and the models belong to each gender, cut-off points are determined as 0,5 for both models since they are providing the maximum rate of accuracy as indicated in Figure 5.11.

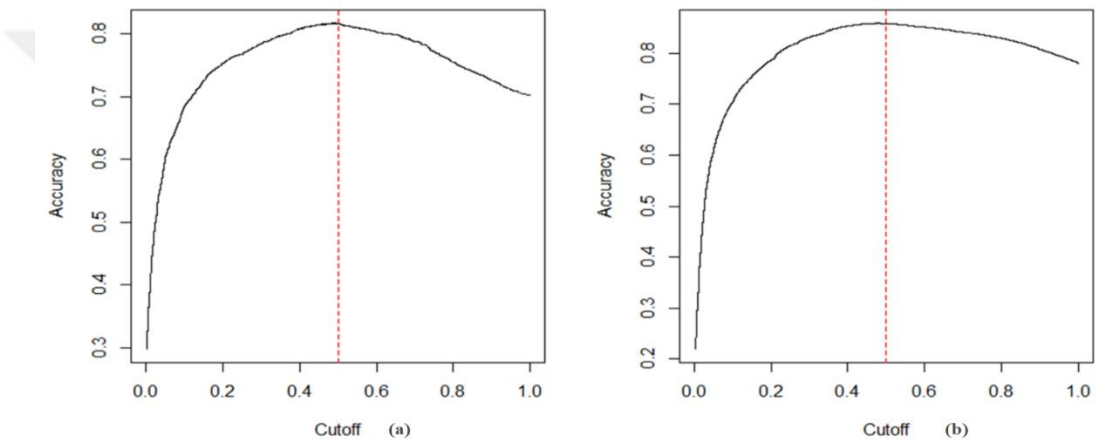

Figure 5.11. Accuracy Rate – Cut-off Point Graph for the Models of East(a) and West(b)

From the perspective of goodness of fit, accuracy rates of the models of East and West are calculated as 81.7% and 85.8%; respectively, which indicate high rate of correct classification. Similarly, ROC curves for the models which are provided in Figure 5.12 also point at important level of goodness of fit with AUC values of 0.88 and 0:90.

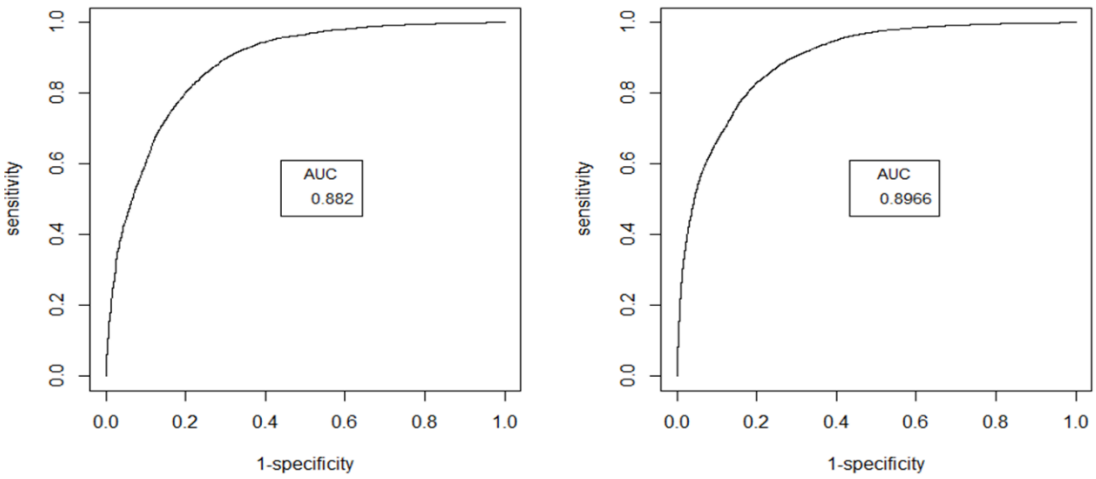

Figure 5.12. ROC Curves for the Models of East and West

Table 5.4 gives the AME's of the micro-factors of unregistered employment for Eastern and Western Regions. Standard errors are given in parenthesis, while base categories for the categorical independent variables are female, higher education and more, divorced, employer, construction, between 11 and 49, full time and TR1 Region.

As indicated in Table 5.4, the marginal effect of gender is higher in eastern regions. While the unregistered probability for men is 4.4% lower on the average than women in western regions, it is 6.6% lower in eastern regions, controlling for the other variables. Similarly, the effect of the level of education is also more determinant in eastern regions. The probabilistic difference between each sequential levels of education is higher in eastern regions. For instance, the difference in probability is 12% between graduates of high school and graduates of higher education and more in eastern regions. The same difference in western regions is 7% on the average. Moreover, the difference in the probability of unregistered employment between graduates of primary school and high school is 13.2% and 7.8% on the average for eastern and western regions; respectively.

Table 5.4. AME's of the Micro-Factors of Unregistered Employment for Eastern and Western Regions

	East	West
Gender (male)	-0.0663*** (0.0064)	-0.0440*** (0.0026)
Age	-0.0255*** (0.0012)	-0.0285*** (0.0006)
Age-squared	0.0003*** (0.0000)	0.0004*** (0.0000)
Education (non-educated)	0.4068*** (0.0117)	0.2805*** (0.0079)
Education (primary school)	0.2518*** (0.0077)	0.1474*** (0.0032)
Education (secondary school)	0.1893*** (0.0075)	0.1075*** (0.0033)
Education (high school)	0.1197*** (0.0082)	0.0695*** (0.0038)
Education (vocational high school)	0.0841*** (0.0091)	0.0554*** (0.0035)
Marital Status (married)	-0.0469. (0.0200)	-0.0518*** (0.0058)
Marital Status (never married)	-0.0251 (0.0210)	-0.0341*** (0.0066)
Marital Status (widow)	0.0419 (0.0352)	-0.0188 (0.0121)
Employment Status (own account worker)	0.1624*** (0.0105)	0.1227*** (0.0044)
Employment Status (regular employee)	0.0719*** (0.0094)	0.0476*** (0.0038)
Employment Status (unpaid family worker)	0.2186*** (0.0192)	0.2383*** (0.0082)
Sector (manufacturing)	-0.0154. (0.0083)	-0.0292*** (0.0040)
Sector (service)	-0.0660 (0.0064)	-0.0636*** (0.0035)
Number of Employees (less than 11)	0.2703*** (0.0067)	0.1940*** (0.0032)
Number of Employees (more than 49)	-0.0886*** (0.0060)	-0.0636*** (0.0026)
Type of Employment (part time)	0.1680*** (0.0120)	0.2678*** (0.0051)

* p<0.05, ** p<0.01, *** p<0.001

Regarding the marital status, most remarkably, the probability difference between married and widow individuals in eastern regions is almost twice the corresponding difference in eastern regions. Apart from the widow category, the effects of marital status categories are close to each other. On the other hand, the effect of employment status is slightly higher in eastern regions. Besides, the effects of sectors are close to each other in eastern and western regions.

The size of the workplace is more determinant in eastern regions compared to the western regions. While the probability difference in terms of unregistered employment between an individual working in a workplace with less than 11 employees and individual working in a workplace with more than 49 employees is almost 36% in eastern regions, it is almost 26% in western regions, controlling for the other variables. Conversely, the effect of type of employment is larger in western regions. Namely, being a part time employee increases the probability of unregistered employment by 26.8% on the average compared to being a full-time employee in western regions while it increases 16.8% on the average in eastern regions.

5.5.3.1. Revealing the Differences among Regional Groups

Group comparisons between regional groups are conducted from the perspective of education level and sectors at the same time. Within this context, profiles with similar specificities in terms of other independent variables are determined in line with the research questions. Accordingly, the predicted probabilities of unregistered employment for persons which are male, 35 years old, married and working as full time, regular employee in a workplace with less than 11 employees are calculated and compared. The z-values for the observation of significance of the difference in probabilities are calculated accordingly.

Table 5.5. Probabilities of Unregistered Employment for Different Regions by Different Levels of Education and Different Sectors

Education Level		Probability of Unregistered Employment		Difference of Probability	Z-value
		East	West		
Service	Non-educated	0.6599	0.4170	0.2429	11.57
	Primary school	0.4213	0.2142	0.2071	17.65
	Secondary School	0.3254	0.1632	0.1622	14.31
	High School	0.2270	0.1201	0.1069	9.90
	Vocational High School	0.1815	0.1056	0.0759	6.86
	Higher Education and More	0.0892	0.0570	0.0322	5.39
Manufacturing	Non-educated	0.7431	0.5007	0.2424	11.43
	Primary school	0.5204	0.2765	0.2440	15.07
	Secondary School	0.4183	0.2147	0.2036	12.21
	High School	0.3045	0.1606	0.1439	8.67
	Vocational High School	0.2484	0.1420	0.1064	6.49
	Higher Education and More	0.1274	0.0781	0.0493	6.04
Construction	Non-educated	0.7649	0.5664	0.1985	10.02
	Primary school	0.5497	0.3324	0.2174	14.71
	Secondary School	0.4472	0.2626	0.1846	11.65
	High School	0.3300	0.1995	0.1305	7.85
	Vocational High School	0.2710	0.1773	0.0937	5.54
	Higher Education and More	0.1411	0.0994	0.0417	4.02

When the Table 5.5 is analyzed, the regional disparity in terms of unregistered employment is obvious. The difference in probability of unregistered employment is significant and in favour of the western regions for every level of education and every sector. With minor exceptions, the probabilities of unregistered employment decreases with the increasing level of education. For instance, the difference of unregistered employment probability in service sector is almost 24.3% between non-educated individuals in eastern and western regions, while it is 3.2% for graduates of higher education and more. Similarly, the difference is 24.2% for non-educated individuals in manufacturing sector, while it is 10.6% and 4.9% for graduates of vocational high schools and graduates of higher education and more; respectively. Also, it should be noted that, the difference in probability apparently increases when the level of education increases from non-educated to primary school graduation in construction sector. This causes from the fact that the decrease in probability with the corresponding increase in education level in eastern regions cannot reach the same decrease in western regions unlike other sectors and education levels.

It is inferred from the Table 5.5 that the regional disparities is remarkable for every main sectors in Turkey. However, the level of education is a critical determinant regarding its magnitude similar to gender disparity in terms of unregistered employment.

CHAPTER 6

CONCLUSION

The objective of the thesis is to conduct a comprehensive individual based micro-level analysis of unregistered employment in Turkey. In this respect, logistic regression models are established to detect the determinants of an individual's engagement in unregistered employment, different models for different groups including genders, regions and years are compared and econometric analysis depending on the results are performed.

In order to accomplish the objective of the study, making the comparisons of logistic regression models for different groups is a key step to be taken. However, group comparison of logistic regression models in a similar way with OLS is manipulating depending on the unobserved heterogeneity in logistic regression. In this sense, this study focuses on the group comparison problem in logistic regression.

In order to get to the root of the comparison problem in logistic regression, the theoretical background of the logistic regression is explained with the latent propensity interpretation in which the extent of the dependent variable's closeness to success is taken into consideration. In this respect, the discussions on the diagnosis and the remediation of the problem in the literature are revealed and analyzed.

Considering that the tests of differences in predicted probabilities based on the models and the marginal effects are not scaled by unobserved heterogeneity unlike model coefficients, the comparisons between gender, region and year groups in terms of unregistered employment are conducted by means of predicted probabilities and marginal effects. Differently from marginal effects, testing differences in predicted probabilities requires to define a profile in order to control the independent variables other than the one which is focus of interest. In this respect, the profile of a person,

whose probability of unregistered employment is being measured, is determined depending on the research question. On the other hand, the application and interpretation of marginal effects are more straightforward and general. In this respect marginal effects are used in comparison of all groups including region, gender and year. On the other hand, using the test of difference in predicted probabilities for comparison is illustrated in region and gender groups.

The models within the scope of the thesis are built with the microdata provided by Turkstat's HLFS. Basically, there are 9 different independent variables taken into consideration to measure their effect on unregistered employment including gender, age, education status, sector, number of employees in the workplace, employment status, type of employment and region. However, there are changes in the models exposed to group comparisons. For instance, gender is not an independent variable any more in models built to compare gender groups. On the other hand, the models' goodness of fits are measured by accuracy rates and areas under ROC Curves which indicate outstanding levels of fits in most cases and very close to outstanding in the rest.

The results of the overall model show that the determinants of the unregistered employment are in line with the thematic knowledge, hence expectations. However, the model provides the necessary knowledge regarding the strength of the determinants while comparing the categories to each other. Most remarkably, the model puts forth that being female is associated with an important level of increase in probability of unregistered employment.

Comparison of the gender groups in terms of determinants of unregistered employment indicates various important results. One of the most important results is that the effect of education is greater for women than men. In other words, the decrease in the probability of working unregistered in return of increasing level of education is usually higher for women. Similarly, women's unregistered employment is more sensitive to the sector in terms of unregistered employment. On the other hand, the

marital status is a more powerful determinant for men which may be associated with their patriarchal bread winner role.

The most remarkable result of the comparison of region groups shows that the effect of gender, education, marital status and the number of employees in the workplace are larger in eastern regions. Conversely, the effect of employment type is larger in western regions. Also, the comparison of year groups shows that the marginal effects of the determinants change and most of the differences between the marginal effects of the categories narrow by time. Some of the changes in marginal effects by time are considered to be associated with policy measures.

It should be noted that convenient group comparison methods in logistic regression are limited to compare 2 groups in the literature. In this respect, the groups to be compared in this thesis are selected considering the limitation of two groups in addition to the thematic knowledge and discussions in the field of unregistered employment. Accordingly, the areas in which the comparisons made, are selected among the ones in which there are two groups or more than two groups which can be accumulated within two groups.

In this context, this thesis can be improved if a method to compare more than two groups in logistic regression is developed. Accordingly, further detailed analysis can be made by including comparisons of different groups like sectors and employment status.

In addition, a future work can be addressed to check the presence of unobserved heterogeneity problem of group comparison in logistic regression argued in this thesis. For this purpose, a full model including interactions can be built and compared with the results of group analysis.

REFERENCES

Açıkalın, N. (2007). Enformel Sektör ve Yoksulluk: Kentsel İşgücü Pazarı ÜzerineEtkileri İstanbul ve Gaziantep Örnekleri. Sosyoekonomi Journal.

Adam, M.C. & Ginsburgh, V. (1985). The Effects of Irregular Markets on Macroeconomic Policy : Some Estimates for Belgium. European Economic Review, 29(1), 15-33.

Allison, P. D. (1999). Comparing Logit and Probit Coefficients Across Groups. Sociological Methods & Research, 28(2), 186-208.

Allison, P. (2013). Why I don't trust the Hosmer-Lemeshow test for logistic regression. Retrieved el, 19.

Amemiya, T. (1985). Advanced Econometrics. Harvard University Press.

Angel-Urdinola, D. F., & Tanabe, K. (2012). Micro-Determinants of Informal Employment in the Middle East and North Africa Region. World Bank.

Angrist, J. D. (2001). Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice. Journal of Business & Economic Statistics, 19(1), 2-28.

Aydın, E., Hisarcıklılar, M., & İlkkaracan, I. (2010). Formal versus Informal Labor Market Segmentation in Turkey in the course of Market Liberalization. Topics in Middle Eastern and North African Economies, 12.

Başlevent, C., & Acar, A. (2015). Recent Trends in Informal Employment in Turkey. Yildiz Social Science Review, 1(1), 77-88.

Bracha, A., & Burke, M. A. (2014). Informal Work Activity in the United States: Evidence from Survey Responses. Current Policy Perspectives, 14, 1-47.

Buis, M. L. (2016). Logistic Regression: When Can We Do What we Think We Can Do. Unpublished Note, V. 2.3.

Bulutay, T., & Taştı, E. (2004). Informal Sector in the Turkish Labour Market (No. 2004/22). Turkish Economic Association Discussion Paper.

Charmes, J. (1990). A Critical Review of Concepts, Definitons and Studies in the Informal Sector. Turnham/Salome/Schwarz (der.) The Informal Sector Revisited, OECD, Paris.

Chow, G. (1960). Tests of Equality between Sets of Coefficients in Two Linear Regressions. Econometrica, 28:591-605.

Çetintaş, H., & Vergil, H. (2003). Türkiye'de Kayıtdışı Ekonominin Tahmini. Doğuş Üniversitesi Dergisi, Cilt 4, Sayı 1.

DPT (2001). Kayıt Dışı Ekonomi Özel İhtisas Komisyonu Raporu. Sekizinci Beş Yıllık Kalkınma Planı.

Doğrul, H. G. (2012). Determinants of Formal and Informal Sector Employment in the Urban Areas of Turkey. International Journal of Social Sciences and Humanity Studies, 4(2), 217-231.

Eurofound (2009). Measures to Tackle Undeclared Work in the European Union. Eurofound.

Eurofound (2013). Tackling Undeclared Work in Turkey. Eurofound.

European Commission (2007). Stepping Up the Fight Against Undeclared Work. COM(2007) 628 final, Brussels.

Fidan, H., & Genç, S. (2013). Unregistered Employment and Analysis of Factors Affecting the Unregistered Employment: Turkish Private Sector. Mehmet Akif Ersoy Universitesi Sosyal Bilimler Enstitusu, 5(9), 137–150.

Frey, B.S. & Weck-Hanneman, H. (1984) The Hidden Economy as an Unobserved Variable. European Economic Review, 26 (1), 33-53.

Gasparini, L., & Tornarolli, L. (2009). Labor Informality in Latin America and the Caribbean: Patterns and Trends From Household Survey Microdata. Revista Desarrollo y Sociedad, (63), 13-80.

Görmüş, A. (2017). The Micro Determinants of Informal Youth Employment in Turkey. Unregistered Employment, 157-169.

Güloğlu, T. (2005). The Reality of Informal Employment in Turkey. International Programs Visiting Fellow Working Papers, Cornell University.

Hosmer, D.W. & Lemeshow, S. (1980). A Goodness-of-Fit Test for the Multiple Logistic Regression Model. Communications in Statistics A10:1043-1069.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons, 8-37.

ILO (1972). Employment, Incomes and Equity: A Strategy for Increasing Productive Employment in Kenya. International Labour Office, Geneva.

ILO (1991). The Dilemma of the Informal Sector, Report of the Director-General. International Labour Conference, 78th Session. International Labour Office, Geneva.

ILO (2002). Decent Work and the Informal Economy. Report of the Director-General. International Labour Conference, 90th Session; Report VI. International Labour Office, Geneva.

ILO (2013). Labour Inspection and Undeclared Work in the EU. International Labour Office, Labour Administration and Inspection Programme (LAB/ADMIN), Geneva.

ILO (2018). Women and Men in the Informal Economy: A Statistical Picture (Third Edition). International Labour Office – Geneva.

Kan, E. Ö. (2012). Essays on Informality in the Turkish Labor Market. Doctoral Dissertation, Middle East Technical University.

Kılıç, A. (2008). The Gender Dimension of Social Policy Reform in Turkey: Towards Equal Citizenship?. Social Policy & Administration, 42(5), 487-503.

Kuha, J., & Mills, C. (2017). On Group Comparisons with Logistic Regression Models. Sociological Methods & Research, 0049124117747306.

Lehmann, H. (2015). Informal Employment in Transition Countries: Empirical Evidence and Research Challenges. Comparative Economic Studies, 57(1), 1-30.

Levent, H. E., Taştı, E. & Sezer, D. (2004). İşgücü Piyasasının Katmanlı Yapısı. Turkiye'de İşgücü Piyasasının Kurumsal Yapısı ve İşsizlik, pp. 27–63, Tüsiad, İstanbul.

Long, J. S., & Mustillo, S. A. (2018). Using Predictions and Marginal Effects to Compare Groups in Regression Models for Binary Outcomes. Sociological Methods & Research, 0049124118799374.

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences, 7.

Long, J. Scott. (2009). Group Comparisons in Logit and Probit Using Predicted Probabilities. Department of Sociology, University of Indiana.

Maddala, G. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Mc Fadden, (1974). The Measurement of Urban Travel Demand. Journal of Public Economics, 3: 303-328.

Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It. European sociological review, 26(1), 67-82.

Radchenko, N. (2014). Heterogeneity in Informal Salaried Employment: Evidence from the Egyptian Labor Market Survey. World Development,V62, 169-188

Sahoo, B. K., & Neog, B. J. (2017). Heterogeneity and Participation in Informal Employment among Non-Cultivator Workers in India. International Review of Applied Economics, 31(4), 437-467.

Sarılı, M. A. (2004). Türkiye'de Kayıt Dışı Ekonominin Boyutları, Nedenleri, Etkileri Ve Alınması Gereken Tedbirler. Bankacılar Dergisi, 41.

Schneider, F. (2000) The Growth of the Shadow Economy in the OECD: Some Preliminary Explanations. Journal of International Affairs, 53 (2), 413-431.

Tansel, A., & Kan, E. O. (2012). Labor Mobility across the Formal/Informal Divide in Turkey: Evidence from Individual Level Data. IZA Discussion Papers No. 6271.

Tutz, G. (2012). Regression for categorical data(Vol. 34). Cambridge University Press, 71.

Tang, W., He, H., & Tu, X. (2012). Applied categorical and count data analysis. Chapman and Hall/CRC, 150-152.

Van Eyck, K. (2003). Flexibilizing Employment: An Overwiew. Small Enterprise Development Job Creation and Enterprise Department (SEED.), Working Paper No. 41, International Labour Office.

Williams, C. C., & Horodnic, I. A. (2018). Extent and Distribution of Unregistered Employment in the Service Industries in Europe. The Service Industries Journal, 38(11-12), 856-874.

Williams, C. C., & Kayaoğlu, A. (2017). Evaluating the Prevalence of Employees Without Written Terms of Employment in the European Union. Employee Relations, 39(4), 487-502.

Williams, R. (2009). Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups. Sociological Methods & Research, 37(4), 531-559.

Windebank, J., & Horodnic, I. A. (2017). Explaining Participation in Undeclared Work in France: Lessons for Policy Evaluation. International Journal of Sociology and Social Policy, 37(3/4), 203-217.

World Bank (2010) Turkey Country Economic Memorandum – Informality: Causes, Consequences, Policies, World Bank, Washington DC.

Yavuz, A. (1995). Esnek Çalışma ve Endüstri İlişkilerine Etkisi. Filiz Kitabevi.

APPENDICES

A. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for 2017

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.589e+00	1.347e-01	11.794	< 2e-16
Gender (male)	-4.487e-01	2.156e-02	-20.812	< 2e-16
Age	-2.688e-01	5.140e-03	-52.290	< 2e-16
Agesq	3.511e-03	6.001e-05	58.508	< 2e-16
Education (non-educated)	2.518e+00	4.813e-02	52.314	< 2e-16
Education (primary school)	1.599e+00	3.174e-02	50.376	< 2e-16
Education (secondary school)	1.261e+00	3.279e-02	38.467	< 2e-16
Education (high school)	8.502e-01	3.734e-02	22.768	< 2e-16
Education (vocational high school)	7.221e-01	3.791e-02	19.046	< 2e-16
Marital status (married)	-5.000e-01	4.887e-02	-10.232	< 2e-16
Marital status (never married)	-3.281e-01	5.540e-02	-5.922	3.17e-09
Marital status (widow)	-8.200e-02	1.018e-01	-0.805	0.4207
Employment status (own account worker)	1.185e+00	4.078e-02	29.044	< 2e-16
Employment status (regular employee)	5.360e-01	3.859e-02	13.890	< 2e-16
Employment status (unpaid family worker)	1.997e+00	6.017e-02	33.189	< 2e-16
Sector (manufacturing)	-2.643e-01	3.132e-02	-8.441	< 2e-16
Sector (service)	-6.192e-01	2.640e-02	-23.455	< 2e-16
Number of employees (less than 11)	1.695e+00	2.638e-02	64.253	< 2e-16
Number of employees (more than 49)	-1.094e+00	3.911e-02	-27.975	< 2e-16
Type of employment (part time)	1.965e+00	3.263e-02	60.200	< 2e-16
Region (TR2)	2.432e-01	4.112e-02	5.914	3.34e-09
Region (TR3)	-1.770e-01	3.572e-02	-4.956	7.20e-07
Region (TR4)	-3.702e-02	3.956e-02	-0.936	0.3493
Region (TR5)	8.501e-02	3.561e-02	2.387	0.0170
Region (TR6)	3.196e-01	3.543e-02	9.019	< 2e-16
Region (TR7)	-8.653e-02	4.450e-02	-1.944	0.0518
Region (TR8)	6.179e-02	4.069e-02	1.519	0.1289
Region (TR9)	-2.937e-01	5.483e-02	-5.356	8.49e-08
Region (TRA)	3.564e-01	4.924e-02	7.237	4.59e-13
Region (TRB)	7.777e-01	4.314e-02	18.027	< 2e-16
Region (TRC)	8.689e-01	3.787e-02	22.945	< 2e-16

Note: Base category for the independent variables: female, higher education and more, divorced, employer, construction, between 11 and 49, full time and TR1 Region.

Null deviance: 138337 on 126822 degrees of freedom
Residual deviance: 85037 on 126792 degrees of freedom
AIC: 85099

Number of Fisher Scoring iterations: 6

B. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for 2007

	Estimate	Std.Error	zvalue	Pr(>|z|)
(Intercept)	1.662e+00	1.521e-01	10.925	<2e-16
Gender (male)	-4.170e-01	2.584e-02	-16.140	<2e-16
Age	-1.933e-01	5.332e-03	-36.261	<2e-16
Agesq	2.308e-03	6.476e-05	35.641	<2e-16
Education (non-educated)	2.803e+00	5.876e-02	47.702	<2e-16
Education (primary school)	1.845e+00	3.817e-02	48.326	<2e-16
Education (secondary school)	1.493e+00	4.202e-02	35.531	<2e-16
Education (high school)	9.037e-01	4.231e-02	21.361	<2e-16
Education (vocational high school)	8.100e-01	4.661e-02	17.378	<2e-16
Marital status (married)	-5.617e-01	8.154e-02	-6.888	5.65e-12
Marital status (never married)	-1.179e-01	8.501e-02	-1.386	0.165595
Marital status (widow)	-3.071e-01	1.287e-01	-2.386	0.017017
Employment status (own account worker)	1.013e+00	4.053e-02	24.987	<2e-16
Employment status (regular employee)	1.344e+00	3.956e-02	33.978	<2e-16
Employment status (unpaid family worker)	2.646e+00	6.278e-02	42.150	<2e-16
Sector (manufacturing)	-1.107e+00	3.668e-02	-30.185	<2e-16
Sector (service)	-1.317e+00	3.340e-02	-39.436	<2e-16
Number of employees (less than 10)	1.594e+00	2.535e-02	62.870	<2e-16
Number of employees (more than 49)	-1.302e+00	3.400e-02	-38.294	<2e-16
Type of employment (part time)	1.486e+00	7.010e-02	21.194	<2e-16
Region (TR2)	-2.940e-01	4.551e-02	-6.462	1.04e-10
Region (TR3)	-5.669e-01	3.066e-02	-18.494	<2e-16
Region (TR4)	-3.810e-01	3.399e-02	-11.209	<2e-16
Region (TR5)	-4.168e-01	3.687e-02	-11.306	<2e-16
Region (TR6)	1.216e-01	3.393e-02	3.583	0.000339
Region (TR7)	-3.161e-01	4.833e-02	-6.540	6.14e-11
Region (TR8)	-4.246e-01	3.600e-02	-11.792	<2e-16
Region (TR9)	-4.448e-01	5.193e-02	-8.566	<2e-16
Region (TRA)	3.054e-02	5.512e-02	0.554	0.579542
Region (TRB)	3.707e-01	4.855e-02	7.636	2.24e-14
Region (TRC)	8.109e-01	4.090e-02	19.827	<2e-16

Notes: Base category for the independent variables: female, higher education and more, divorced, employer, construction, between 11 and 49, full time and TR1 Region.

Null deviance: 119296 on 93332 degrees of freedom
Residual deviance: 79701 on 93302 degrees of freedom
AIC: 79763

Number of Fisher Scoring iterations: 5

C. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Males

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	2.106e+00	1.606e-01	13.115	<2e-16
Age	-2.848e-01	5.855e-03	-48.646	<2e-16
Agesq	3.706e-03	6.739e-05	54.993	<2e-16
Education (non-educated)	2.081e+00	6.120e-02	34.003	<2e-16
Education (primary school)	1.259e+00	3.754e-02	33.538	<2e-16
Education (secondary school)	9.906e-01	3.838e-02	25.813	<2e-16
Education (high school)	6.924e-01	4.411e-02	15.698	<2e-16
Education (vocational high school)	5.672e-01	4.463e-02	12.707	<2e-16
Marital status (married)	-5.937e-01	7.256e-02	-8.182	2.79e-16
Marital status (never married)	-3.604e-01	7.881e-02	-4.573	4.81e-06
Marital status (widow)	-4.432e-01	1.776e-01	-2.496	0.0126
Employment status (own account worker)	1.076e+00	4.293e-02	25.071	<2e-16
Employment status (regular employee)	5.007e-01	4.056e-02	12.345	<2e-16
Employment status (unpaid family worker)	2.218e+00	8.030e-02	27.620	<2e-16
Sector (manufacturing)	-4.341e-01	3.329e-02	-13.039	<2e-16
Sector (service)	-6.905e-01	2.712e-02	-25.463	<2e-16
Number of employees (less than 11)	1.532e+00	3.031e-02	50.525	<2e-16
Number of employees (more than 49)	-1.088e+00	4.509e-02	-24.141	<2e-16
Type of employment (part time)	2.057e+00	4.602e-02	44.686	<2e-16
Region (TR2)	2.082e-01	4.835e-02	4.306	1.66e-05
Region (TR3)	-2.695e-01	4.241e-02	-6.354	2.10e-10
Region (TR4)	-8.606e-02	4.673e-02	-1.842	0.0655
Region (TR5)	-2.375e-02	4.175e-02	-0.569	0.5695
Region (TR6)	2.139e-01	4.173e-02	5.126	2.95e-07
Region (TR7)	-3.216e-01	5.223e-02	-6.157	7.42e-10
Region (TR8)	-8.905e-02	4.813e-02	-1.850	0.0643
Region (TR9)	-4.671e-01	6.399e-02	-7.299	2.89e-13
Region (TRA)	2.250e-01	5.544e-02	4.057	4.96e-05
Region (TRB)	6.515e-01	4.835e-02	13.474	<2e-16
Region (TRC)	7.758e-01	4.266e-02	18.186	<2e-16

Notes: Base category for the independent variables: higher education and more, divorced, employer, construction, between 11 and 49, full time and TR1 Region.

Null deviance: 96940 on 92012 degrees of freedom
Residual deviance: 63565 on 91983 degrees of freedom
AIC: 63625

Number of Fisher Scoring iterations: 6

D. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Females

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.4159975	0.3353629	-4.222	2.42e-05
Age	-0.2298707	0.0117555	-19.554	<2e-16
Agesq	0.0030052	0.0001441	20.849	<2e-16
Education (non-educated)	3.1405650	0.0846045	37.121	<2e-16
Education (primary school)	2.3162619	0.0615155	37.653	<2e-16
Education (secondary school)	1.7592169	0.0660140	26.649	<2e-16
Education (high school)	1.0222959	0.0718223	14.234	<2e-16
Education (vocational high school)	0.8843369	0.0736711	12.004	<2e-16
Marital status (married)	-0.3015639	0.0705094	-4.277	1.89e-05
Marital status (never married)	-0.3154905	0.0881087	-3.581	0.000343
Marital status (widow)	0.0318239	0.1327234	0.240	0.810504
Employment status (own account worker)	1.8528563	0.1416937	13.076	<2e-16
Employment status (regular employee)	0.5955741	0.1336196	4.457	8.30e-06
Employment status (unpaid family worker)	1.6496220	0.1470657	11.217	<2e-16
Sector (manufacturing)	1.1394697	0.1796587	6.342	2.26e-10
Sector (service)	0.5728444	0.1743417	3.286	0.001017
Number of employees (less than 11)	2.0554533	0.0544093	37.778	<2e-16
Number of employees (more than 49)	-1.0824232	0.0780996	-13.860	<2e-16
Type of employment (part time)	1.7869314	0.0519462	34.400	<2e-16
Region (TR2)	0.2640258	0.0801660	3.293	0.000990
Region (TR3)	0.0058010	0.0685731	0.085	0.932582
Region (TR4)	-0.0247120	0.0766788	-0.322	0.747241
Region (TR5)	0.3376572	0.0714577	4.725	2.30e-06
Region (TR6)	0.5182933	0.0695038	7.457	8.85e-14
Region (TR7)	0.5391389	0.0923579	5.837	5.30e-09
Region (TR8)	0.3515279	0.0795254	4.420	9.86e-06
Region (TR9)	0.1460239	0.1100600	1.327	0.184586
Region (TRA)	0.6401265	0.1142671	5.602	2.12e-08
Region (TRB)	1.0897159	0.1019745	10.686	<2e-16
Region (TRC)	1.1519861	0.0886560	12.994	<2e-16

Notes: Base category for the independent variables: higher education and more, divorced, employer, construction, between 11 and 49, full time and TR1 Region.

Null deviance: 40974 on 34809 degrees of freedom
Residual deviance: 20204 on 34780 degrees of freedom
AIC: 20264

Number of Fisher Scoring iterations: 6

E. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Eastern Regions

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	0.9158554	0.2912693	3.144	0.00166
Gender (male)	-0.5158461	0.0493494	-10.453	<2e-16
Age	-0.2028748	0.0101970	-19.896	<2e-16
Agesq	0.0023244	0.0001202	19.344	<2e-16
Education (non-educated)	2.9861207	0.0879126	33.967	<2e-16
Education (primary school)	2.0060767	0.0710095	28.251	<2e-16
Education (secondary school)	1.5944413	0.0697992	22.843	<2e-16
Education (high school)	1.0981741	0.0770842	14.246	<2e-16
Education (vocational high school)	0.8169814	0.0862741	9.470	<2e-16
Marital status (married)	-0.3671588	0.1535447	-2.391	0.01679
Marital status (never married)	-0.1941218	0.1613561	-1.203	0.22895
Marital status (widow)	0.3186521	0.2667648	1.195	0.23228
Employment status (own account worker)	1.2691263	0.0873310	14.532	<2e-16
Employment status (regular employee)	0.5945924	0.0826523	7.194	6.3e-13
Employment status (unpaid family worker)	1.6731075	0.1440706	11.613	<2e-16
Sector (manufacturing)	-0.1178019	0.0634866	-1.856	0.06352
Sector (service)	-0.5169724	0.0494917	-10.446	<2e-16
Number of employees (less than 10)	1.8193372	0.0528126	34.449	<2e-16
Number of employees (more than 49)	-1.1837243	0.0854702	-13.850	<2e-16
Type of employment (part time)	1.2629396	0.0884567	14.277	<2e-16

Notes: Base category for the independent variables: female, higher education and more, divorced, employer, construction, between 11 and 49 and full time.

Null deviance: 31207 on 25638 degrees of freedom
Residual deviance: 19867 on 25619 degrees of freedom
AIC: 19907

Number of Fisher Scoring iterations: 6

F. Logit Model Estimates of the Micro-Determinants of the Unregistered Employment for Western Regions

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.858e+00	1.508e-01	12.321	<2e-16
Gender (male)	-4.217e-01	2.395e-02	-17.608	<2e-16
Age	-2.825e-01	5.906e-03	-47.838	<2e-16
Agesq	3.751e-03	6.865e-05	54.646	<2e-16
Education (non-educated)	2.471e+00	5.886e-02	41.987	<2e-16
Education (primary school)	1.507e+00	3.556e-02	42.375	<2e-16
Education (secondary school)	1.172e+00	3.737e-02	31.355	<2e-16
Education (high school)	8.149e-01	4.299e-02	18.957	<2e-16
Education (vocational high school)	6.698e-01	4.220e-02	15.874	<2e-16
Marital status (married)	-4.881e-01	5.166e-02	-9.448	<2e-16
Marital status (never married)	-3.142e-01	6.004e-02	-5.233	1.67e-07
Marital status (widow)	-1.699e-01	1.107e-01	-1.535	0.125
Employment status (own account worker)	1.176e+00	4.604e-02	25.535	<2e-16
Employment status (regular employee)	5.100e-01	4.359e-02	11.699	<2e-16
Employment status (unpaid family worker)	2.044e+00	6.608e-02	30.931	<2e-16
Sector (manufacturing)	-2.644e-01	3.628e-02	-7.288	3.15e-13
Sector (service)	-6.021e-01	3.130e-02	-19.234	<2e-16
Number of employees (less than 10)	1.630e+00	3.039e-02	53.651	<2e-16
Number of employees (more than 49)	-1.053e+00	4.406e-02	-23.891	<2e-16
Type of employment (part time)	2.064e+00	3.498e-02	58.992	<2e-16

Notes: Base category for the independent variables: female, higher education and more, divorced, emp loyer,
construction, between 11 and 49 and full time.

Null deviance: 106466 on 101183 degrees of freedom
Residual deviance: 65379 on 101164 degrees of freedom
AIC: 65419

Number of Fisher Scoring iterations: 6

G. R Codes Used in the Analysis

```
#building models

nonagriculturalmodel2017 <- glm(socialsecurity ~ ., family = "binomial", data =
nonagriculturaldata2017)
summary(nonagriculturalmodel2017)

nonagriculturalmodel2007 <- glm(socialsecurity ~ ., family = "binomial", data =
nonagriculturaldata2007)
summary(nonagriculturalmodel2007)

malenonagriculturalmodel2017 <- glm(socialsecurity ~ ., family = "binomial", data
= malenonagriculturaldata2017)
summary(malenonagriculturalmodel2017)

femalenonagriculturalmodel2017 <- glm(socialsecurity ~ ., family = "binomial",
data = femalenonagriculturaldata2017)
summary(femalenonagriculturalmodel2017)

eastnonagriculturalmodel2017 <- glm(socialsecurity ~ ., family = "binomial", data =
eastnonagriculturaldata2017)
summary(eastnonagriculturalmodel2017)

westnonagriculturalmodel2017 <- glm(socialsecurity ~ ., family = "binomial", data =
westnonagriculturaldata2017)
summary(westnonagriculturalmodel2017)

#calculating accuracy rates and area under ROC Curves, drawing cut-off point
#graphes

preds2017 <- predict(nonagriculturalmodel2017, nonagriculturaldata2017)
preds2017 <- exp(preds2017)/(1+exp(preds2017))
preds2017 <- ifelse(preds2017> 0.5, 1, 0)
accuracy2017<- table (preds2017, nonagriculturaldata2017$socialsecurity)
accuracy2017
accuracy2017<-(accuracy2017[1,1]+accuracy2017[2,2])/sum(accuracy2017)
accuracy2017

install.packages("ROCR")
library(ROCR)
install.packages("gplots")
library(gplots)
```

```
pred2017<-predict(nonagriculturalmodel2017, nonagriculturaldata2017, type=
'response')
pred2017<-prediction(pred2017, nonagriculturaldata2017$socialsecurity)
eval2017<-performance(pred2017, "acc")
plot(eval2017)
abline(v=0.5, col="red", type="l", lty=2)

roc2017<-performance(pred2017, "tpr", "fpr")
plot(roc2017)
abline(a=0, b=1)
plot(roc2017, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
auc2017<- performance (pred2017, "auc")
auc2017<-unlist(slot(auc2017, "y.values"))
auc2017
auc2017<-round(auc2017,4)
plot(roc2017, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
legend(.8, .2, auc2017, title="AUC", cex=0.9, xjust=0.9, yjust=0.9)

preds2007 <- predict(nonagriculturalmodel2007, nonagriculturaldata2007)
preds2007 <- exp(preds2007)/(1+exp(preds2007))
preds2007 <- ifelse(preds2007> 0.5, 1, 0)
accuracy2007<- table (preds2007, nonagriculturaldata2007$socialsecurity)
accuracy2007
accuracy2007<-(accuracy2007[1,1]+accuracy2007[2,2])/sum(accuracy2007)
accuracy2007

pred2007<-predict(nonagriculturalmodel2007, nonagriculturaldata2007, type=
'response')
pred2007<-prediction(pred2007, nonagriculturaldata2007$socialsecurity)
eval2007<-performance(pred2007, "acc")
plot(eval2007)
abline(v=0.5, col="red", type="l", lty=2)

roc2007<-performance(pred2007, "tpr", "fpr")
plot(roc2007)
abline(a=0, b=1)
plot(roc2007, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
auc2007<- performance (pred2007, "auc")
auc2007<-unlist(slot(auc2007, "y.values"))
auc2007
auc2007<-round(auc2007,4)
plot(roc2007, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
legend(.8, .2, auc2007, title="AUC", cex=0.9, xjust=0.9, yjust=0.9)
```

```
preds2017male <- predict(malenonagriculturalmodel2017,
malenonagriculturaldata2017)
preds2017male <- exp(preds2017male)/(1+exp(preds2017male))
preds2017male <- ifelse(preds2017male> 0.5, 1, 0)
accuracy2017male<- table (preds2017male,
malenonagriculturaldata2017$socialsecurity)
accuracy2017male
accuracy2017male<-
(accuracy2017male[1,1]+accuracy2017male[2,2])/sum(accuracy2017male)
accuracy2017male

pred2017male<-predict(malenonagriculturalmodel2017,
malenonagriculturaldata2017, type= 'response')
pred2017male<-prediction(pred2017male,
malenonagriculturaldata2017$socialsecurity)
eval2017male<-performance(pred2017male, "acc")
plot(eval2017male)
abline(v=0.5, col="red", type="l", lty=2)

roc2017male<-performance(pred2017male, "tpr", "fpr")
plot(roc2017male)
abline(a=0, b=1)
plot(roc2017male, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
auc2017male<- performance (pred2017male, "auc")
auc2017male<-unlist(slot(auc2017male, "y.values"))
auc2017male
auc2017male<-round(auc2017male,4)
plot(roc2017male, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
legend(.5, .5, auc2017male, title="AUC", cex=0.9, xjust=0.3, yjust=0.3)

preds2017female <- predict(femalenonagriculturalmodel2017,
femalenonagriculturaldata2017)
preds2017female <- exp(preds2017female)/(1+exp(preds2017female))
preds2017female <- ifelse(preds2017female> 0.5, 1, 0)
accuracy2017female<- table (preds2017female,
femalenonagriculturaldata2017$socialsecurity)
accuracy2017female
accuracy2017female<-
(accuracy2017female[1,1]+accuracy2017female[2,2])/sum(accuracy2017female)
accuracy2017female

pred2017female<-predict(femalenonagriculturalmodel2017,
femalenonagriculturaldata2017, type= 'response')
```

```
pred2017female<-prediction(pred2017female,
femalenonagriculturaldata2017$socialsecurity)
eval2017female<-performance(pred2017female, "acc")
plot(eval2017female)
abline(v=0.5, col="red", type="l", lty=2)

roc2017female<-performance(pred2017female, "tpr", "fpr")
plot(roc2017female)
abline(a=0, b=1)
plot(roc2017female, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
auc2017female<- performance (pred2017female, "auc")
auc2017female<-unlist(slot(auc2017female, "y.values"))
auc2017female
auc2017female<-round(auc2017female,4)
plot(roc2017female, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
legend(.5, .5, auc2017female, title="AUC", cex=0.9, xjust=0.3, yjust=0.3)

preds2017east <- predict(eastnonagriculturalmodel2017,
eastnonagriculturaldata2017)
preds2017east <- exp(preds2017east)/(1+exp(preds2017east))
preds2017east <- ifelse(preds2017east> 0.5, 1, 0)
accuracy2017east<- table (preds2017east,
eastnonagriculturaldata2017$socialsecurity)
accuracy2017east
accuracy2017east<-
(accuracy2017east[1,1]+accuracy2017east[2,2])/sum(accuracy2017east)
accuracy2017east

pred2017east<-predict(eastnonagriculturalmodel2017, eastnonagriculturaldata2017,
type= 'response')

par(mfrow=c(1,2))
pred2017east<-prediction(pred2017east, eastnonagriculturaldata2017$socialsecurity)
eval2017east<-performance(pred2017east, "acc")
plot(eval2017east)
abline(v=0.5, col="red", type="l", lty=2)

roc2017east<-performance(pred2017east, "tpr", "fpr")
plot(roc2017east)
abline(a=0, b=1)
plot(roc2017east, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
auc2017east<- performance (pred2017east, "auc")
auc2017east<-unlist(slot(auc2017east, "y.values"))
auc2017east
```

```
auc2017east<-round(auc2017east,4)
plot(roc2017east, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
legend(.5, .5, auc2017east, title="AUC", cex=0.9, xjust=0.3, yjust=0.3)

preds2017west <- predict(westnonagriculturalmodel2017,
westnonagriculturaldata2017)
preds2017west <- exp(preds2017west)/(1+exp(preds2017west))
preds2017west <- ifelse(preds2017west> 0.5, 1, 0)
accuracy2017west<- table (preds2017west,
westnonagriculturaldata2017$socialsecurity)
accuracy2017west
accuracy2017west<-
(accuracy2017west[1,1]+accuracy2017west[2,2])/sum(accuracy2017west)
accuracy2017west

pred2017west<-predict(westnonagriculturalmodel2017,
westnonagriculturaldata2017, type= 'response')
library(ROCR)
library(gplots)
pred2017west<-prediction(pred2017west,
westnonagriculturaldata2017$socialsecurity)
eval2017west<-performance(pred2017west, "acc")
plot(eval2017west)
abline(v=0.5, col="red", type="l", lty=2)

roc2017west<-performance(pred2017west, "tpr", "fpr")
plot(roc2017west)
abline(a=0, b=1)
plot(roc2017west, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
auc2017west<- performance (pred2017west, "auc")
auc2017west<-unlist(slot(auc2017west, "y.values"))
auc2017west
auc2017west<-round(auc2017west,4)
plot(roc2017west, main="ROC Curve", ylab="sensitivity", xlab= "1-specificity")
legend(.5, .5, auc2017west, title="AUC", cex=0.9, xjust=0.3, yjust=0.3)

#calculating average marginal effects

install.packages("margins")
library("margins")
mm.2017<-margins(nonagriculturalmodel2017)
summary(mm.2017)

mm.2007<-margins(nonagriculturalmodel2007)
```

```
summary(mm.2007)

mm.male<-margins(malenonagriculturalmodel2017)
summary(mm.male)

mm.female<-margins(malenonagriculturalmodel2017)
summary(mm.female)

mm.east<-margins(eastnonagriculturalmodel2017)
summary(mm.east)

mm.west<-margins(westnonagriculturalmodel2017)
summary(mm.west)

#drawing confidence interval graphes for differences of probabilities of
#unregistered employment between male and female throughout different levels
#of education (codes for only comparison are exhibited as a sample.)

X<-data.frame("region"=c("TR1"), "age"=c(15), "agesq"=c(225),
"education"=c("primaryschool"),
          "maritalstatus"=c("married"), "employmentstatus"=c("regularemployee"),
          "sector"=c("service"), "numberofemployees"=c("lessthan11"),
          "typeofemployment"=c("fulltime"))

ex <- matrix(c(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 15, 225, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0,
1, 1, 0, 0),30,1)

for(i in 1:65){
  X[2]<-i+14
  X[3]<-(i+14)^2
  ex[13]<-i+14
  ex[14]<-(i+14)^2

  XBmale[i] <- predict(malenonagriculturalmodel2017,X)
  phiXmale <- exp(XBmale)/(1+exp(XBmale))

  XBfemale[i] <- predict(femalenonagriculturalmodel2017,X)
  phiXfemale <- exp(XBfemale)/(1+exp(XBfemale))
}
phiXmale
phiXfemale

age<-data.frame("age"=(15:79))
phioverage<-(cbind(age,phiXmale, phiXfemale))
```

```
phioverage

par(mfrow=c(1,1))

with(phioverage, plot(x=phioverage$age,y=phioverage$phiXmale, type="l",
                    ylab="Prob. of Unregistered Emp. for Primary School Graduates",
xlab= "Age", xlim=c(15, 65), ylim=c(0,1)))
with(phioverage, lines(x=phioverage$age,y=phioverage$phiXfemale, col="red",
type="l",lwd=3))
with(phioverage, lines(x=phioverage$age,y=phioverage$phiXmale,
col="dodgerblue1", type="l",lwd=3))
legend(x=50, y=.2, legend=c("male","female"), fill=c("red", "dodgerblue1"), cex=1)
axis(side=1, at=c(15,65))

upper.c1 <- numeric(65);
lower.c1 <- numeric(65)

for(i in 1:65){
  X[2]<-i+14
  X[3]<-(i+14)^2
  ex[13]<-i+14
  ex[14]<-(i+14)^2

  XBmale[i] <- predict(malenonagriculturalmodel2017,X)
  phiXmale <- exp(XBmale)/(1+exp(XBmale))
  fXBmale <- exp(-XBmale)/(1+exp(-XBmale))^2
  varBhatmale<-vcov(malenonagriculturalmodel2017)
  varphimale[i]<-fXBmale[i]*t(ex)%*%varBhatmale%*%ex*fXBmale[i]

  XBfemale[i] <- predict(femalenonagriculturalmodel2017,X)
  phiXfemale <- exp(XBfemale)/(1+exp(XBfemale))
  fXBfemale <- exp(-XBfemale)/(1+exp(-XBfemale))^2
  varBhatfemale<-vcov(femalenonagriculturalmodel2017)
  varphifemale[i]<-fXBfemale[i]*t(ex)%*%varBhatfemale%*%ex*fXBfemale[i]

  differenceofprobabilities<-phiXfemale-phiXmale

  upper.c1<- (phiXfemale-phiXmale)+1.96*sqrt(varphifemale+varphimale)
  lower.c1<- (phiXfemale-phiXmale)-1.96*sqrt(varphifemale+varphimale)
}
varphimale
varphifemale

differenceofprobabilities
```

```
upper.c1
lower.c1
age<-data.frame("age"=(15:79))

confidenceinterval<-cbind(age, upper.c1, lower.c1, differenceofprobabilities)
confidenceinterval

with(confidenceinterval, plot(x=confidenceinterval$age,
y=confidenceinterval$differenceofprobabilities, type="l",
                        ylab="Difference in Probability for Primary School Graduates ",
                        xlab="Age", xlim=c(15, 65), ylim=c(-0.25,0.25),
xaxt="n",yaxt="n"))
axis(side=1, at=c(15,25,35,45,55,65))
axis(side=2, at=c(0, 0.05, 0.10, 0.15, 0.20),cex.axis=.8, las=1)
abline(h=c(0.5, 0.10, 0.15, 0.20), lty=3)
abline(h=c(0), lty=1)

with(confidenceinterval,
lines(confidenceinterval$age,confidenceinterval$differenceofprobabilities,,
col="dodgerblue1", type="l",lwd=3))
with(confidenceinterval, lines(confidenceinterval$age,confidenceinterval$upper.c1,
lty=5, col="RED",type="l",lwd=1))
with(confidenceinterval, lines(confidenceinterval$age,confidenceinterval$lower.c1,
lty=5, col="RED",type="l",lwd=1))

polygon(c(confidenceinterval$age,rev(confidenceinterval$age)),c(confidenceinterval
$upper.c1, rev(confidenceinterval$lower.c1)),
      col="gray", density = c(50, 50), border=NA)

#calculating Z- Statistic for the probability differences in unregistered
#employment between different regions for different sectors and education
#levels (codes for one comparison are exhibited as a sample.)

X<-data.frame("gender"=c("male"), "age"=c(35), "agesq"=c(1225),
"education"=c("noneducated"),
          "maritalstatus"=c("married"), "employmentstatus"=c("regularemployee"),
"sector"=c("service"),
          "numberofemployees"=c("lessthan11"), "typeofemployment"=c("fulltime"))
X

XBeast<-predict(eastnonagriculturalmodel2017,X)
XBeast

phiXeast<- exp(XBeast)/(1+ exp(XBeast))
```

```
phiXeast

fXBeast <- exp(-XBeast)/(1+exp(-XBeast))^2
fXBeast

varBhateast <- vcov(eastnonagriculturalmodel2017)
varBhateast

ex <- matrix(c(1, 1, 35, 1225, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0),20,1)

varphieast<-fXBeast*t(ex)%*%varBhateast%*%ex*fXBeast
varphieast

XBwest<-predict(westnonagriculturalmodel2017,X)
XBwest

phiXwest<- exp(XBwest)/(1+ exp(XBwest))
phiXwest

fXBwest <- exp(-XBwest)/(1+exp(-XBwest))^2
fXBwest

varBhatwest <- vcov(westnonagriculturalmodel2017)
varBhatwest

varphiwest<-fXBwest*t(ex)%*%varBhatwest%*%ex*fXBwest
varphiwest

zteststat=(phiXeast-phiXwest)/sqrt(varphieast+varphiwest)
zteststat
```