Three Essays on the Informal Sector

by

Musah Khalid

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Economics

University of Manitoba

Winnipeg, Manitoba

Thesis Advisors: Dr. John Serieux Author: Musah Khalid
Dr. Ryan Godwin, Dr. Mahmoud Torabi

Three Essays on the Informal Sector

ABSTRACT

This thesis consists of three essays that examine: heterogeneity in informal wage employment; access and use of health insurance by individuals in the informal sector; and the gender wage gap in the sector.

The first essay uses data from the sixth round of the Ghana Living Standards Survey (GLSS) and employs a finite mixture model to analyze the structure of informal wage employment, and test for evidence of segmentation in the Ghanaian labour market. The findings show that Ghanaian informal wage employment comprises two divisions –higher-paid and lower-paid– each with a distinct wage function. This result, we find, is robust to two definitions of informal wage employment –non-payment of employment income tax and non-coverage of social security. The sizes of the two divisions is, however, sensitive to how informality is defined. Assuming workers are earnings-maximizers, we find evidence of segmentation in the labour market implying that pervasive barriers to labour mobility may be preventing some workers from working in the division of the (labour) market where their potential earnings will be maximized. More specifically, we find that 72% to 79% of all informal wage workers are involuntary employed in the sector, and would be better off in other divisions of the labour market. In addition, using propensity score matching, we estimate that

an average worker in informal wage employment, who could maximize their potential earnings in the formal sector, is potentially losing on the average ($108.6–$148.7) monthly due to entry barriers or inadequate employment opportunities in the formal sector.

The second essay uses propensity score matching to analyze how different ways of acquiring health insurance affects health care utilization. We take advantage of a natural experiment occasioned by the introduction of the Ghana National Health Insurance Scheme in 2003. Even though the scheme mandates every adult to buy health insurance this (mandate) is not enforced outside the formal sector. The result is that formal workers get automatically enrolled and their 'contributions' deducted from their earnings while everyone else may voluntarily buy health insurance. Using data from the World Bank's Living Standards Measurement Survey (LSMS) we examine the determinants of voluntary health insurance uptake and how this impacts health care use. We find robust evidence that older, wealthier and more educated individuals are more likely to buy health insurance, and that the average individual who voluntarily enrolls in the program uses more health care (about 35%) compared to an identical formal sector worker who is involuntary enrolled.

The third essay addresses two questions: what is the relative magnitude of the gender wage gap across the conditional wage distribution of formal and informal sector wage earners, and what explains the gap in each sector. This essay also uses data from the sixth round of the Ghana Living Standards Survey (GLSS) and employs a quantile regression technique to estimate the wage functions and decompose the wage gap across the 25^{th}, 50^{th} and 75^{th} quantiles. In addition, we attempt to correct for the non-random participation of women in the labour market using an identification-at-infinity technique. In the informal sector, we find a 44% wage gap at the 25^{th} quantile which increases to 52% at the 75^{th} quantile. This contrasts with 5%–10%

between same quantiles in the formal sector. After correcting for selection, however, the gap increases at all quantiles in the informal sector, but turns positive at the lower to median quantiles in the formal sector. This general pattern suggest significantly higher levels of discrimination in the informal sector, implying that the sector may be a cause of poverty for some female workers employed in the sector.

ACKNOWLEDGMENTS

I am deeply grateful to my thesis advisor, Professor John Serieux, for his invaluable support throughout my years at the University of Manitoba. From my first day at the university when I knocked on his door to ask for a research assistantship job till today he has always been there to help. I have benefited so much from our discussions and debates. Without his encouragement and constant support, I could not have completed my thesis and doctoral studies. I am also highly indebted to Professors Ryan Godwin and Mahmoud Torabi for their excellent advice and insightful comments, and discussions which greatly improved this thesis. My deepest gratitude also goes to Professor John Loxley for his financial support and guidance throughout this journey. I would also not have made it without that support.

I will also like to acknowledge Professor Albert Sumell of Youngstown State University, whose advising and support made this journey a success. My sincere appreciation also goes to the staff and faculty of the Department of Economics at the University of Manitoba, especially Alan Nabess, Betty McGregor, Professors Pinaki Bose, Julia Witt, Umut Oguzoglu, Fletcher Baragar, Ian Hudson, Ryan Compton and Wayne Simpson for the various support they gave me during the course of my program. Finally, I thank my student colleagues, especially Mehdi Arzandeh, whose penetrating inquiry helped shaped and brought clarity to some of my ideas.

DEDICATION

I thank my mom, Hasia, for her prayers and support. To my late dad, I say I made it, as you had predicted. Loads of appreciation also to my uncles Ishak Alhassan, Baba Kumasi, Mustapha and Awudu for their investment in my education without which I wouldn't reach this point. My grandaunt, Habiba Adama, probably deserves more credit than anyone else for raising me to be the person I am today. To my former high school headmistress, Mrs. Aba Tachie-Menson, who pushed me to strive for excellence, I say this is for you.

Most importantly, I thank my wife, Mariam Senusi, and my two children, Khalida and Ishak, for their love and patience especially on the days when I had to spend additional time working on this thesis. This thesis is dedicated to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The term informal sector was coined by Hart [1973] while studying the urban labour market of Ghana. He observed that the labor market of Accra comprised a "modern sector", made up enterprises that run with some measure of bureaucracy and are amenable to documentation, and a traditional sector that employs the "reserve army of the underemployed and unemployed", and whose activities are difficult to measure because the government lacks the institutional capacity to do so. He further emphasized the potential productive value of the sector, and cautioned that the central idea of economic development as shifting labour from that sector to the formal economy may be counter productive.

Even though Hart [1973] coined the term 'informal sector', it wasn't until the International Labour Organization [ILO, 1972] used it to describe the labour market of Kenya that it would gain widespread popularity. In 2002, the ILO estimated that the sector accounted for 72% of all non-agricultural employment in sub-Saharan Africa, 71% in Asia and 51% in Latin America [ILO, 2002]. According to the most recent estimates, in 2014 the sector accounted for 82% of total non-agricultural employment in South Asia, 66% in sub-Saharan Africa, 65% in East and South-East Asia (excluding China) and 51% in Latin America [ILO, 2013]. The marginal change in the relative size of the sector, especially in sub-Saharan Africa, signifies the persistence of the phenomenon of informality. Given the typically lower average wages in the sector, the phenomenon of informality has a major implication for poverty alleviation. For policymakers to get it right they may need a thorough and complete understanding of the dynamics of the sector, and how it interacts with other parts of the economy.

Through three essays, this thesis uses more recent datasets to analyze the nature and composition of wage employment in the sector, the dynamics of the (informal) gender wage gap, and the differential impact of informality on health care use.

While the term 'informal sector' may have been popularized by Hart [1973] and the ILO [1972], the theoretical framework of dual or segmented labour market had preceded that. The theory of dual labour market has its origin in the works of Lewis [1954], Harris and Todaro [1970], Doeringer and Piore [1985], Stiglitz [1974] among others. The central idea of the theory is that labour markets, especially in developing countries, can be distinguished by the presence of two distinct sectors –primary and secondary– characterized by within-sector, but limited inter-sector mobility. Workers in the primary sector (formal sector) are said to enjoy relatively high wages with stable employment and good working conditions while those in the secondary sector (informal sector) are consigned to low-paid jobs with unstable tenure and poor working conditions. Without imperfect sectoral mobility, and inadequate jobs in the primary sector, workers in the secondary sector will move to the primary sector where returns to their characteristics will be relatively better rewarded. In this reading, therefore, informality is a result of structural or institutional constraints and informal employment is considered involuntary.

However, more recently, Maloney [2004], Loayza and Rigolini [2006] and others have argued that the informal sector is instead composed of individuals who voluntarily choose to work in that sector because it offers them the highest net benefit. Following from this, Fields [1990] hypothesized that the sector may comprise two divisions –upper and lower-tier. The upper-tier is said to contain workers who voluntarily choose to work in the sector, and the lower-tier workers rationed out of the formal sector. Therefore, the informal sector may actually contain both voluntary and involuntary employment types. The first essay tests for heterogeneity in informal wage employment, and the degree of segmentation in the Ghanaian labour market assuming workers are earnings-maximizers using data from the sixth round of the Ghana Living Standards Survey (GLSS VI) and a finite mixture model. The findings show that the Ghanaian labour market is segmented, and informal wage employment comprises two divisions; a higher-paid and lower-paid division. But, in contrast to the Fields hypothesis, we find that both the higher-paid and lower-paid divisions contain voluntary and involuntary employment types. Thus, not only do we reject the claim that all informal work is voluntary, we also do not find evidence to support the theory that each informal division contains only voluntary or involuntary workers.

The second essay attempts to provide empirical evidence on how voluntary health insurance uptake impacts the use of health resources. The essay takes advantage of a natural experiment occasioned by the establishment of Ghana's National Health Insurance Scheme (NHIS) in 2003. In spite of the fact that the program mandates every adult to purchase health insurance, the government has been unable to enforce that rule due to lack of capacity. As a result, formal workers are automatically enrolled in the scheme and 'contributions' deducted directly from their employment income while informal members have to voluntarily enroll by paying annual income-adjusted premiums. Two issues arise out of this arrangement. First, the health insurance (theoretical) literature has shown that in the presence of voluntary health insurance sicker people are more likely to buy health insurance [Rothschild and Stiglitz, 1976]. Second, a possible loss of coverage due to the inability to renew health insurance could drive up health care use because of loss aversion [Kőszegi and Rabin, 2006]. Using data from the World Bank's Living Standards Measurement Survey (LSMS) we explore the determinants of voluntary health insurance uptake among informal members, and use propensity score matching to examine the difference between the health care use of an informal member (of the scheme) and an identical formal sector worker. We find robust evidence that older, richer and sicker individuals are more likely to buy health insurance, and that voluntary health insurance causes a higher rate of use of health care resources.

The last essay examines the gender wage gap along the conditional wage function in the informal sector. Given that the sector makes up more than 50% of the labour market of the majority of African countries, the existence of a relatively large gender wage gap in the sector will necessarily have major implications for poverty reduction especially among women. While much has been written about the existence and magnitude of the (formal sector) gender wag gap in many countries [1] very little is known about the gap in the informal sector. Given that the informal sector is the largest sector of majority of African economies, a study of the wage gap in Africa would be incomplete without an analysis of the nature of the gap in the informal sector. This essay fills this gap in the literature by comparing the manner in which the gap varies along the conditional wage function in the informal sector to that in the formal sector using quantile regression and data from the sixth Ghana Living Standard Survey (GLSS VI). In addition, we correct for sample-selection using an identification-at-infinity strategy. To the best of our knowledge this is the first study of the gender wage gap in the Ghanaian setting using quantile regression with correction for selection. The results indicate an increasing wage gap along the conditional wage distribution of the informal sector. In the formal sector, however, after correction for self-selection, the wage gap is positive at the lower to median quantiles but the advantage disappears at the upper quantiles. Therefore, as expected, compared to the informal sector women are better off in the formal sector.

[1] See [Weichselbaumer and Winter-Ebmer, 2005] for a detailed analysis of the wage gap in different countries.

2. HETEROGENEITY IN INFORMAL WAGE EMPLOYMENT IN GHANA: COMPETITIVE ADVANTAGE OR LAST RESORT

2.1 Introduction

Labour markets in developing countries, up until recently, have traditionally been modeled as comprising a high productive formal sector with attractive wages and working conditions, and a low productivity informal sector that absorbs workers who lack access to the formal sector. Wages in the formal sector, due to institutional or efficiency-wage reasons, are said to be set above market clearing levels which, therefore, creates a barrier to entry into formal employment for a section of the working population [Fields, 1990; Harris and Todaro, 1970; Lewis, 1954; Stiglitz, 1976]. In this regard, the informal sector is considered a sector of last resort for those who seek to avoid unemployment. Workers in the informal sector are said to earn less relative to identical formal sector workers, and in the absence of entry barriers, informal workers would enter the formal sector. This view of labour markets implies that informal sector employment is involuntary, and observed differences in wages of identical workers in the two sectors is a consequence of segmentation.

However, recent empirical studies have cast doubt on the validity of this hypothesis. Some studies have found that a significant proportion of employment in the informal sector may actually reflect deliberate choices of individuals given their skills endowments, preferences and earnings potential in the sector [Maloney, 2004; Saavedra and Chong, 1999; Yamada, 1996]. In this reading, informal employment does not comprise disadvantaged workers rationed out of the formal sector, but rather workers who choose to work in the sector based on comparative advantage reasons or the existence of more desirable non-wage characteristics in the sector [Gindling, 1991; Maloney, 2004]. Fields [1990] argues that, indeed, these two informal labour types may actually coexist in the sector. Specifically, he posits that the sector consists of two distinct divisions –upper-tier and a lower-tier. The upper-tier is said to be competitive and comprise largely of individuals who, given their endowments, expect to earn more than they would in the formal sector. The lower-tier, on the other hand, represents the involuntary part of the sector and is made up of individuals rationed out of the formal sector.

Focusing on identifying heterogeneity in informal wage employment this essay follows the framework of Günther and Launov [2012], who use data from Cote D'Ivoire and employ a finite mixture model with sample selection to detect clusters within informal employment based on differences in earning functions. Their findings show that Cote D'Ivoire's informal sector is composed of two distinct divisions. In addition, they show that the distribution of workers across the entire labour market is inefficient if it is assumed that workers are earnings-maximizers, implying a segmented labour market.

The contribution of this paper is twofold. First, using data from the sixth round of the Ghana Living Standards Survey (GLSS) we test the hypothesis of heterogeneity in informal wage employment in the Ghanaian labour market, and attempt to identify the sizes and earning functions in each division. In addition, assuming workers are earning-maximizers, we test the hypothesis of segmentation in the entire labour market. Essentially, we compare the counterfactual distribution of workers across divisions of the entire labour market (when it is assumed that workers are earning-maximizers) with the actual empirical distribution of workers. A significant difference in the two distributions imply some sort of segmentation in the Ghanaian labour market. Second, we use propensity score matching to estimate the informal wage gain and penalty. The informal wage gain is defined as the wage differential between a typical worker in informal wage employment who maximizes their earnings in any of the informal divisions and an identical formal worker. The wage penalty, on the other hand, is defined as the wage differential between a typical worker in informal wage employment who could maximize their earnings in the formal sector and an identical formal worker. To the best of our knowledge, this essay is the first empirical study that examines the heterogeneous structure of informal wage employment in Ghana, and estimates the sizes of the various divisions that comprise the sector and their corresponding wage gaps; providing a case study of the labour market of one of Africa's lower-middle income economies.

The findings support the dualistic view of informal wage employment comprising higher-paid and lower-paid divisions even though their sizes are sensitive to how informal wage employment is defined. In addition, we find that each division of informal wage employment is composed of both voluntary and involuntary employment in contrast to the Fields-hypothesis that the higher-paid and lower-paid divisions comprise voluntary and involuntary workers, respectively. Our estimates also indicate a significant average monthly wage penalty of \$108.6–\$148.7, but we find no significant wage gain meaning there is no significant difference between the earnings of a typical worker in voluntary informal wage employment and an identical formal worker.

The remainder of the paper is organized as follows. The next section reviews the theoretical and empirical literature on the theory of dual or segmented labour markets. Section 2.3 describes the data, variable selection and identification of informal wage employment. We present the econometric methodology in section 2.4. Estimation of results are discussed in section 2.5 and section 2.6 concludes.

2.2 Literature Review

The theory of segmented labour markets has its foundations in the works of Lewis [1954], Harris and Todaro [1970], Doeringer and Piore [1985], and Stiglitz [1974] among others. Two main ideas are central to segmented labour market theory. First, that the labour market can be thought of as comprising several, but mainly two – formal and informal– distinct sectors each with a different wage setting rule. One sector (formal) is said to be higher-paid where workers enjoy relatively better working conditions and job security while the other sector (informal) is lower-paid with limited job tenure and worse working conditions. Second, access to jobs in the higher-paid sector is rationed due to limited availability of jobs or some institutional factors leading to job queues. The queuing for jobs by the workers in the informal sector implies that informal work is involuntary. This assertion has been challenged by some researchers [eg Loayza and Rigolini, 2006; Maloney, 2004] who have argued that workers move freely between the formal and informal sectors, and they choose the sector that offers them the highest net benefit based on their preferences and skills set. Subsequently, Fields [1990] has attempted to integrate the two views by hypothesizing that the informal sector may actually contain two distinct divisions –voluntary higher-paid and involuntary lower-paid divisions.

Several empirical studies have tested the validity of the labour market segmentation hypothesis. The central idea of the test is to look for clustering effects and a sign of non-integrated components. Labour markets are said to be competitive if no significant clustering effects are found. The results, however, have been mixed, with a number of studies supporting or rejecting the hypothesis of segmentation. Most studies of informality in Latin America have validated the competitive labour market hypothesis. For example, Magnac [1991] tested for segmentation in the labour market of Columbia using a generalized Roy model. The results indicate that the Columbian labour market is competitive. Similar results were found by Pratap and Quintin [2006] who employ a semi-parametric approach to study the labour market of Argentina, and Gindling [1991] who uses a generalized regression model with selection correction to test segmentation in the labour market of Costa Rica. On the other hand, studies by Dickens and Lang [1985], Tannuri-Pianto and Pianto [2002], and Heckman and Hotz [1986] have all found evidence in favour of labour market segmentation. A major shortcoming of these studies is the a priori assumption of a homogeneous informal sector, which has been challenged by Fields [1990, 2004], Perry et al. [2007] and Paulson and Townsend [2005]. As a result, more recent studies, especially on Africa, have explored the hypothesis of a heterogeneous informal labour market comprising both voluntary and involuntary divisions.

Using household data from five African countries, Dimova et al. [2010] study heterogeneity in the informal sector of five African cities –Cotonou, Ouagadougou, Abidjan,Bamako, Dakar– defining informal divisions based on self-employment *vs* wage employment. They found evidence of segmentation suggesting inefficient allocation of labour across the entire labour market. Günther and Launov [2012] and Salem and Bensidoun [2012] both employ a finite mixture model to study the labour markets of Cote D'Ivoire and Egypt, respectively, assuming workers are earnings-maximizers. They show that the distribution of workers across the labour market is inefficient, and that the informal sector contains two divisions with each comprising both voluntary and involuntary employment. More recently, Radchenko [2014] employs a model of essential heterogeneity, which allows for both the exploration of differences in observable and unobservable skills of formal and informal workers, and uses Egyptian panel data from 1998-2006. The results support both the segmentation and competitive market hypothesis. Specifically, she finds evidence of a heterogeneous informal sector and inefficient allocation of labour in the entire labour market in 1998. But by

2006, the dynamics have changed with the informal sector containing both voluntary and involuntary labour types, an indication of labour market segmentation.

2.3 *Data*

This paper uses data from the sixth round of the Ghana Living Standards Survey (GLSS) conducted in 2012/13. The survey is a nationally representative survey designed to collect information on the general living conditions in Ghana. The GLSS VI surveyed 18,000 households, collecting individual and household information on socioeconomic topics such as education, health, employment, income and housing conditions. This round was selected because it is the most recent and comprehensive survey on the labour force of Ghana.

The formal sector is defined as: all economic activities –public or private– that falls under state regulation. We define formal employees as workers covered by social security. Informal wage employment, on the other hand, is defined to comprise (non-agricultural) wage workers without access to social security. Thus, we follow the legalistic definition of informal wage employment proposed by the International Labour Organization [ILO, 1993]. This definition characterizes informal wage workers as those who do not pay taxes on employment income, are without a written employment contract or have no social security coverage [Hussmanns, 2004]. To ensure that the results are robust to how informality is defined we compare the results with one obtained using the non-payment of employment income definition of informality.

From the survey, the working-age population (persons aged 15–65) is estimated at 79.6%. Among this, 74.4% are employed, 5.2% unemployed and 20.4% inactive. A majority of the employed (52.4%) are in self-employment, 20.2% in wage employment, 22.3% are family workers and the remaining are apprentice and casual workers. Our analysis covers workers between the ages of 15–65 and in non-agricultural wage employment. This comprises 16% of all employment in the labour market. Because this study focuses on wage employment, self-employed individuals are dropped. Also, workers in either formal or informal wage employment who report no wages (1204 individuals)[1] are also dropped leaving a final sample of 4,480 individuals. Using the non-coverage of social security definition of informal wage employment this sample divides into 1,730 and 2750 formal and informal workers, respectively. The composition of the sample changes to 2,211 and 2,269, respectively, if the non-payment of taxes on employment income definition of informality is employed. Therefore, on the net, some 481 formerly informal individuals become formal workers by this definition. A possible explanation of this composition change may be that the cost of social security compliance may outweigh that of tax compliance making employers more likely to pay employment taxes than extend social security coverage to workers. As it turns out, this sample size difference has significant impact on the findings as we will see later. The sample used for estimating the selection model includes the 4,480 individuals in non-agricultural wage employment with positive observed earnings and 5,522 inactive (which includes the unemployed). The earnings measure used is monthly wage.

Figure 2.1 below presents the kernel density plot of the log of monthly earnings for formal and informal workers. The estimated mean earnings is reported in Table 2.1. The average monthly wage of a typical informal wage worker is 42% lower than that of a formal worker. The log monthly wage distribution of informal workers shows less dispersion compared to that of formal workers, indicating the relative variability in formal wages. The story is a bit different if the non-payment of employment in-

[1] We are aware of the consequences this might have on our results, but we proceed on the basis that there is no systematic differences between these workers and those with observed positive wages.

come taxes definition of informality (see A.1 in appendix) is used. The variability in log monthly wage of formal workers falls, but increases for informal workers. Even though the two definitions imply slightly different log wage distributions, overall both (2.1) and (A.1) show some overlap between the formal and informal density functions, indicating that not all informal wage employment pay less than formal jobs.

Figure 2.1: Kernel density of monthly earnings

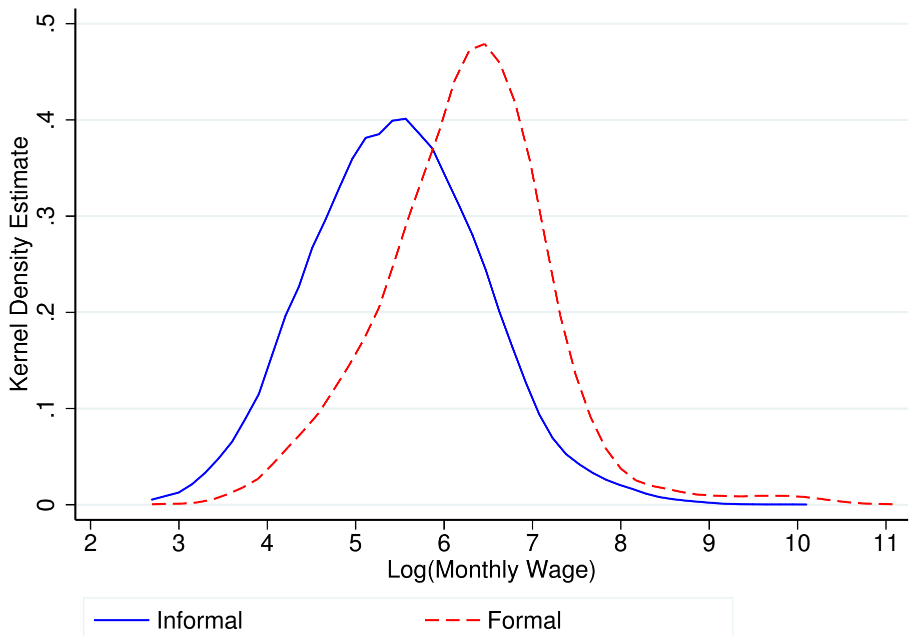


Table 2.1 below presents the means and standard deviations of variables used in the study population. Unsurprisingly, educational levels are highest in the formal sector, with 74% of formal workers having completed more than ten (10) years of education. This compares to 33% of all workers in informal wage employment and 31% of the inactive population. With regards to age, more than 80% of the inactive population are between the ages of 15–24 compared to 8% and 22% formal and informal workers, respectively. This large disparity could be due to the fact that the inactive population comprise a disproportionate number of students. Additionally, both formal and informal wage employment are male and urban biased. Specifically, 76% of formal

Table 2.1: Descriptive statistics of the labour market

	Inactive		Formal		Informal	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
Monthly wage			927.444	2366.870	386.145	616.178
Education						
None	0.058	0.234	0.024	0.154	0.085	0.279
Low (1-6 years)	0.178	0.383	0.034	0.180	0.136	0.343
Medium (7-10 years)	0.453	0.498	0.208	0.406	0.453	0.498
High (>10 years)	0.311	0.463	0.734	0.442	0.325	0.469
Age						
15-24	0.803	0.398	0.083	0.275	0.218	0.413
25-44	0.137	0.343	0.620	0.486	0.609	0.488
Age $\geq$ 44	0.060	0.238	0.298	0.457	0.173	0.379
Industry						
Mining			0.034	0.180	0.073	0.259
Manufacturing			0.079	0.269	0.105	0.307
Construction			0.034	0.182	0.111	0.314
Commerce			0.080	0.272	0.229	0.420
Transportaion & Storage & Communication			0.072	0.258	0.189	0.392
Other						
Location (Urban=1)	0.631	0.483	0.758	0.429	0.712	0.453
Sex (Male=1)	0.465	0.499	0.655	0.475	0.693	0.461
Marital status (Married=1)	0.109	0.312	0.668	0.471	0.567	0.496
Exclusion variables in selection equation						
Household headship	0.059	0.236	0.714	0.452	0.639	0.480
Household size	6.011	3.113	4.092	2.494	4.122	2.617
Children in HH	1.702	1.768	1.236	1.384	1.315	1.462
Infants in HH	0.080	0.284	0.097	0.305	0.102	0.312
HH members aged $\geq$ 60	0.316	0.575	0.114	0.364	0.147	0.415
N	5522		1730		2750	

workers and 71% of informal workers live in urban areas compared to 63% in the active population. As expected, only 11% of the inactive population is married. This compares to 67% and 57% of formal and informal workers, respectively. There are also marked differences in the distribution of workers across industries. More than 70% of informal workers are employed in five industries –mining, manufacturing, commerce, construction and transportation, storage and communication– compared to 30% of formal workers. This implies that a majority of formal and informal workers are concentrated in different industries. The analysis, however, focuses on these (five) industries because they provide sufficient overlap in the distribution of workers between the two sectors to allow for meaningful comparison. In that regard, our base category for comparison becomes other industries.

Additional variables used in the selection equation includes: a dummy for household headship, household size, number of children in the household (defined as children under the age of 14), dummy for whether there are infants (children under the age of 1) in household, and a dummy for whether there are individuals 60 years and older in household. These variables are selected because they affect the individual's opportunity cost of non-participation in the labour market, but do not directly impact their earnings potential. The summary statistics show systematic differences in the distribution of these variables between individuals out of the labour market and those in wage employment. More specifically, only 6% of inactive individuals are head of their households compared to 71% and 64% of formal and informal workers, respectively. Similarly, inactive individuals come from larger households compared to individuals in employment. Overall, compared to the working population, the average inactive individual comes from a larger household with large proportions of both young and older individuals. These systematic differences are likely to impact differently the opportunity cost of participation in wage employment which could lead to possible self-selection into work.

2.4 Econometric Methodology

2.4.1 Finite Mixture Model

We assume the entire labour market is made up two sectors –formal and informal– but the informal sector can be heterogeneous. That is, we assume the formal sector is a single distinct division but the informal sector can comprise a number of divisions with regards to wage employment. From the data, we observe whether an individual is inactive or whether they belong to the formal or informal sector. We, however, do not observe affiliation to any of the heterogeneous informal divisions. As such, we cannot determine which proportion of informal wage employment is voluntary, and which is involuntary. One way of solving this identification problem is to use the distribution of observed wages to determine the number of latent divisions within the (informal) sector. This is because if there are unique divisions in the sector then each will be distinguished by its own distinct wage function.

Following the framework of Günther and Launov [2012], the log-earnings for the entire labour market, Y, consisting of j divisions Y_j such that $Y = \bigcup_{j=1}^{j} Y_j$ can be given as;

$$\ln Y_{ij} = x_i' \beta_j + \epsilon_{ij}, \quad i \in Y_j, \quad \epsilon_{ij} \sim N(0, \sigma_j^2) \tag{2.1}$$

where Y_j are the earnings of an individual i in division j. The error term (ϵ) is normally distributed with a mean of zero and a variance, (σ_j^2), and errors are uncorrelated across divisions. This implies that the earnings distribution in each division is distinct and independent, and the returns (β_j) to individual characteristics (x_i) varies from division to division.

Given the non-random participation of individuals in the labour market, the estimation of the wage function in each division may suffer from selection bias [Heckman, 1979]. To account for this potential selection bias, we assume that the individual's decision to work is a function of a set of personal characteristics (z_i) as follows:

$$y_{is} = z_i'\gamma + \epsilon_{is}, \quad \epsilon_{is} \sim N(0,1) \tag{2.2}$$

such that the earnings y_{is} is observed only if the outcome of equation (2.2) is positive. However, if the error terms from equations (2.1) and (2.2) are correlated, our estimated (β_j) will be biased. Assuming that the error terms in equations (2.1) and (2.2) follow a bivariate normal distribution with a correlation coefficient, (ρ_j), the distribution of observed wages in the j^{th} division of the labour market is given as follows:

$$f(y_{ij}|y_{is} > 0) = \frac{1}{\sigma_j}\varphi\left(\frac{lny_{ij} - x_i'\beta_j}{\sigma_j\Phi(z_i'\gamma)}\right)\Phi\left(\frac{z_i'\gamma + \frac{\rho_j}{\sigma_j}\left[lny_{ij} - x_i'\beta_j\right]}{\sqrt{1-\rho_j^2}}\right) \tag{2.3}$$

where $\varphi(.)$ and $\Phi(.)$, respectively, represents the density and cumulative density functions of the standard normal distribution. Subsequently, the conditional distribution of wages in the entire labour market can be derived from the distribution of wages in each division using size of each division as weights. However, since affiliation to a division is unobserved we estimate the probability, $P(i \in Y_j) = \pi_j$, that an individual i belongs to division j such that each division is composed of homogeneous workers with regards to the relationship that link their observed wages to individual characteristics. Given this, the conditional distribution of observed wages in the entire labour market is given as;

$$f(y_i) = \sum_{j=1}^{J} \pi_j f(y_i \mid y_{is} > 0, \rho, \sigma_j, \beta_j) \tag{2.4}$$

This model is the finite mixture model with sample selection.

The implementation involves estimating a probit of the participation decision, and deriving estimates of the selection parameters (γ) in the first stage. These parameters are then used as consistent estimates to control for selection in the mixture regression. Given that the proportion of formal employment is known, the mixture is only estimated on the informal sector population.

2.4.2 *Is the labour market segmented?*

Even though results from previous empirical studies have differed on whether labour markets are segmented, they all agree that the existence of a wage gap or different wage setting mechanisms in two sectors do not necessarily imply segmentation. For segmentation to exist, there should be entry barriers or evidence of queuing for jobs in the sector with higher average wages [Dickens and Lang, 1985; Gindling, 1991].

Following from this, we assume that workers are earnings-maximizers who know the wage function in each division, and are able to estimate expected returns to their characteristics in each division of the labour market. If this holds, then we expect that workers will choose to work in the division where expected returns to their observable characteristics is the highest. From this, the counterfactual distribution of workers across divisions can be estimated as:

$$P(i \in Y_j) = P\left(E[lny_{ij} \mid y_{is} > 0; x_i] = \max_{l,l \in [1,J]} E[lny_{il} \mid y_{is} > 0; x_i]\right) \tag{2.5}$$

The segmentation test consist of testing the null hypothesis that the labour market is competitive. Specifically, we test that there are no entry barriers to any division of the labour market, and workers are found in the division that pays them the highest return on their observable characteristics. In practice, this involves testing if the counterfactual distribution of workers across divisions and the actual distribution are the same. Using $\bar{\pi}_j$ to denote the counterfactual distribution of division j the null hypothesis to be tested can be formally stated as:

$$H_0 : \frac{\pi_j}{\bar{\pi}_j} = 1 \tag{2.6}$$

Rejecting the null hypothesis implies the labour market is segmented.

2.4.3 *Formal and informal wage employment: wage differentials*

We also estimate the average wage differentials between voluntary and involuntary informal workers and identical formal counterparts. If, indeed, the labour market is segmented, then we can estimate the informal wage gain –the wage differential between a typical worker in voluntary informal wage employment and an identical formal worker, and the informal wage penalty –the wage differential between a typical worker in involuntary informal wage employment and an identical formal worker. This will be equivalent to estimating the causal impact of voluntary or involuntary informal wage employment on earnings. This estimation problem can be formalized as follows;

$$\tau = E[Y(1) - Y(0) \mid W = 1] \tag{2.7}$$

where, τ, is the average treatment effect on the treated (ATT), defined as the average difference in earnings of voluntary or involuntary informal workers and their potential earnings in the formal sector; $Y(1)$ is the potential earning from voluntary or involuntary informal wage employment; $Y(0)$ is potential earnings of working in the formal sector, and W denotes treatment status, such that $W = 1$ if a worker receives treatment, that is being voluntary or involuntary employed in the informal sector. The basic problem in estimating treatment effects is that only one of the potential outcomes is observed. In our case, we only observe earnings of voluntary or involuntary informal wage employment but not the counterfactual earnings. Rubin [1974] argues that if data can be obtained for a set of potential untreated units with identical potential outcomes, then a comparison can be made between the two groups since, but for the treatment, the (two) groups will be identical.

In this study, the method of propensity score matching is used to construct the counterfactual outcome using outcome of workers in the formal sector. The propensity score, first introduced by Rosenbaum and Rubin [1983], is defined as the conditional probability of receiving treatment conditioned on observable pre-treatment characteristics. Formally this is given as;

$$p(x) \equiv Pr(W = 1 \mid X = x) = E[W \mid X = x] \tag{2.8}$$

Where $W = [0, 1]$ is the indicator of exposure to treatment, and X is the vector of unit pre-treatment characteristics. They show that matching on the propensity score produces consistent estimates of the treatment effect if, conditioned on the propensity scores, potential outcomes are independent of treatment status, and if enough overlap exists between the propensity score distributions of the treated and untreated groups. The former hypothesis, referred to as the conditional independence assumption (CIA) or unconfoundedness, is given as;

$$Y(0), Y(1) \perp W \mid p(x) \tag{2.9}$$

where $\perp$ denotes independence; $Y(0), Y(1)$ represents potential outcomes with no treatment and treatment, respectively. The latter assumption is the overlap or common support condition, given as:

$$0 < P(W = 1 \mid p(x) < 1) \tag{2.10}$$

This hypothesis states that workers with the same propensity scores values, $p(x)$, have a positive probability of being in both the treated and control group [Heckman et al., 1999].

If the propensity score, $p(x)$, is known then the average treatment on the treated (ATT) estimator will be:

$$\tau^{\text{ATT}} \equiv E[E(Y(1) \mid p(x), W = 1) - E(Y(0) \mid p(x), W = 1)] \tag{2.11}$$

$$\tau^{\text{ATT}} \equiv E[Y(1) - Y(0) \mid W, p(x)] \tag{2.12}$$

However, because the distribution of the propensity score is continuous the probability of observing two workers with exactly the same propensity score, $p(x)$, is theoretically zero. As such, techniques like nearest neighbor, kernel and radius matching are commonly employed in estimating the treatment effects[2]. To check the consistency of our estimates we employed all three matching techniques.

2.5 *Empirical Results*

2.5.1 *Optimal informal divisions*

Before we can estimate the mixture model we need to first determine the appropriate number of divisions in the heterogeneous informal wage employment that best fit the data. To do so, we run the mixture with different numbers of a priori imposed informal divisions, and use information criteria to determine which specification best fits the data.

[2] A more detailed analysis of these methods are provided in Caliendo [2006] and Dehejia and Wahba [2002].

Model selection results are presented below in Table 2.2. We see that results from both the AIC and BIC indicate that a two-division informal wage employment is, statistically, more plausible than a homogeneous one. However, the comparison between two and three divisions is ambiguous. While the BIC points to a two-division specification as the best fit, the AIC indicates a three-division model. This is unsurprising given the tendency of the AIC to overfit the data by preferring a larger model, because its penalty for over-parameterization is less stringent compared to the BIC [Kuha, 2004]. In the case of finite mixture models, Cameron and Trivedi [2005] caution against over-parameterization as this poses a risk to identification of the model parameters. We, therefore, choose the parsimonious specification (BIC), and opt for the model with the two informal division. To check whether the number of divisions is sensitive to the definition of informality we run the same test on a sample of informal workers defined as workers who do not pay taxes on employment earnings. The results (see Table A.1 in appendix) confirms two types informal wage employment. Therefore, the number of informal divisions is not sensitive to how informality is defined.

Table 2.2: Information criteria for model selection

Informal sector	AIC	BIC
Homogeneous	40697.48	40780.30
2-division	36849.23	37032.73
3-division	36825.33	37103.54

2.5.2 Estimation results

Tables 2.3 below present the results of the estimation of equation 2.3 – heckman selection regression. The first important finding is the insignificance of the correlation coefficient (ρ). In addition, all household level characteristics, apart from household headship and number of children in household, are insignificant implying no significant differences in the household-level characteristics between workers and inactive individuals. The insignificance of the correlation coefficient (ρ) implies that the unobserved characteristics between observations in and out of the labour force is similar, and therefore no systematic self-selection into the labour market exist. Given this result, we estimate the mixture model without accounting for selection.

Table 2.3: Heckman selection regression

Variable	Coefficient
Constant	-1.601 (-18.77)***
Education *(ref: No school)*	
Low (1-6 years)	0.152 (1.93)
Medium (7-10 years)	0.193 (2.77)*
High (> 10 years)	0.464 (6.64)***
Age *(ref: 15–24)*	
25–44	1.016 (24.22)***
45 $\geq$	0.527 (8.50)***
Other	
Location (Urban=1)	-0.032 (-0.84)
Sex (Male=1)	0.242 (6.86)***
Marital status (Married=1)	0.868 (19.89)***
Household headship	1.365 (27.42)***
Household size	0.007 (0.69)
Children in HH	-0.055 (-3.07)*
Infants in HH	-0.074 (-1.17)
HH members aged $\geq$ 60	-0.058 (-1.71)
ρ	-0.068 (2.07)
N	9993

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Estimates of the mixture and wage regressions (equation 2.4) are presented below in Table 2.4. From the table, we see that Informal-1, the lower-paid informal division, accounts for just 9% of wage employment in the entire labour market, but approximately 14% of all informal wage employment, while Informal-2, the higher-paid informal division, makes up more than 52% of wage employment in the entire labour market, but 86% of all informal wage employment[3]. Additionally, average earnings of formal workers is five times more than the average earnings in the lower-paid informal division, but only two times more than that of the higher-paid informal division. Therefore, earnings are higher in the formal sector compared to both divisions of informal wage employment. There also exist a significant earnings gap between the higher-paid and lower-paid informal divisions. Specifically, average earnings in the lower-paid informal division is only 44% of the earnings in the higher-paid division.

Significant differences also exists in returns to observable characteristics across divisions of the labour market. Returns to all levels of education are highest in the formal sector. However, levels of education doesn't impact wages in the lower-paid informal division, but do matter (especially higher education) in the higher-paid informal division. Similar observation pertains with returns to age. Older workers are better compensated in the formal sector, but age only impacts earnings in the higher-paid informal division. Living in an urban area significantly impacts earnings in all divisions of the labour market though the premium is higher in the informal sector. Specifically, the urban premium for formal workers is only one half of what it is in the lower-paid informal division, but about 92% of that in the higher-paid informal division. This difference in the rural-urban wage gap in the two sectors could possibly

[3]Using non-payment of employment tax definition of informality, however, the higher-paid informal division's proportion of informal wage employment falls to 72%, and the share of total labour market wage employment also falls to 36%. The lower-paid division's share, on the other hand, rises to 14% total labour market wage employment and 28% informal wage employment.(see Table A.2 in appendix). In addition to the changes in composition of employment, total informal wage employment falls from 61% to 51%.

Table 2.4: Mixture model estimation with 2-division informal wage employment

Variable	Formal	Informal-1	Informal-2
Constant	4.322 (36.55)***	3.994 (22.52)***	4.592 (26.29)***
Education *(ref: No school)*			
Low (1-6 years)	1.80 (1.41)	-0.156 (-1.05)	0.209 (1.59)
Medium (7-10 years)	0.375 (3.66)***	0.119 (0.91)	0.295 (2.66)*
High (> 10 years)	1.001 (10.03)***	0.236 (1.64)	0.652 (5.53)***
Age *(ref: 15–24)*			
25–44	0.585 (8.99)***	0.155 (1.92)	0.340 (4.50)***
45 $\geq$	0.798 (11.01)***	0.084 (0.71)	0.445 (4.41)***
Other			
Location (Urban=1)	0.165 (3.80)***	0.315 (4.24)***	0.180 (2.80)*
Sex (Male=1)	0.190 (4.71)***	0.526 (7.13)***	0.551 (8.17)***
Marital status (Married=1)	0.265 (6.51)***	0.133 (1.98)**	0.251 (4.15)***
Industry (main sectors)			
Mining	0.642 (5.34)***	1.475 (10.57)***	0.337 (2.73)*
Manufacturing	-0.130 (-1.87)	0.259 (2.18)**	-0.223 (-2.07)**
Construction	0.096 (1.07)	1.020 (9.19)***	-0.097 (-0.85)
Commerce	-0.262 (-4.23)***	0.243 (2.63)*	-0.263 (-3.11)*
Transportaion & Storage & Communication	0.064 (0.97)	0.496 (4.15)***	-0.169 (-1.71)
π	0.386	0.089	0.525
Average earnings	927.44	184.99	420.08
N	1730	397	2353

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

be explained by the existence of legal and institutional regimes, such as unions, in the formal sector that ensure harmonization of earnings across location.

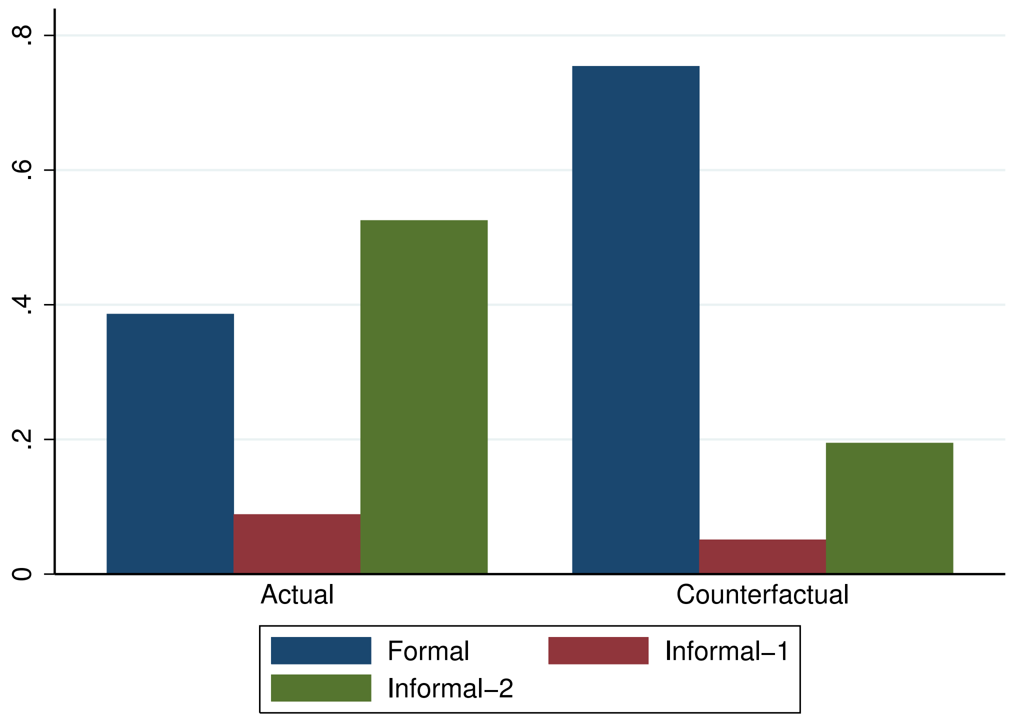
The gender wage gap is also significantly lower in the formal sector compared to both informal divisions. To be concrete, the gap is only about 35% of that in the two informal divisions combined. Possible explanations for this could be that high-skilled women self-select into the formal sector, or there exist effective anti-wage discrimination laws in the formal sector which are not present in the informal sector. There is also systematic differences in earnings by industry. Compared to other industries, returns are significantly higher in the lower-paid informal division than in both the higher-paid informal division and formal sector. In fact, over all, industry, location and gender are the significant determinants of wages in the lower-paid informal division. The general pattern of the results is consistent with results derived from defining informality as non-payment of taxes on employment income (see Table A.2 in appendix).

2.5.3 Segmentation test

Even though significant differences exist in the returns to observable characteristics across divisions, that, in itself, is not an indication of segmentation in the labour market, so long as no entry barriers exist across divisions of the labour market. To test whether the labour market is segmented we assume that workers are earnings-maximizers, and compare the observed distribution of workers across divisions with the counterfactual distribution under the condition of perfect sectoral mobility. The counterfactual distribution is obtained by estimating the proportion of workers in each division for which, conditioned on their observable characteristics, earnings are maximized in that division. If no entry barriers exist then the actual distribution of workers across divisions should be identical to the counterfactual distribution. By contrast, the presence of significant differences between the actual and counterfactual distributions is an indication of some constraints on sectoral mobility. In practice, we test the null hypothesis that the ratio of the actual and counterfactual distributions in each division of the labour market is one. If for any division this ratio is significantly lower than one, then the proportion of workers working in that division is significantly lower than the proportion that would choose to work in that division if no entry barriers existed. Conversely, a ratio greater than one implies the proportion of workers working in that division is significantly higher than the proportion who would voluntarily choose to work in the division.

Figure 2.2 presents the plots of the actual and counterfactual distributions. We can see that the proportion of workers who would be better off in the formal sector is significantly higher than the proportion that work in the sector. Significant differences also exist in the actual and counterfactual distributions of the high-paid (Informal-2) and lower-paid (Informal-1) informal divisions. Indeed, the proportion of workers who could be better off in the higher-paid and lower-paid informal divisions is significantly lower than the proportion currently working in those divisions. This result is robust to the use the non-payment of tax on employment income definition of informality (see figure A.2 in appendix).

Figure 2.2: Distribution of workers across divisions



The result of the segmentation test is presented below in Table 2.5. The result show that, the actual and counterfactual distributions of workers across all three divisions of the labour market are significantly different. The ratios of the actual to the counterfactual distribution is greater than one in both informal divisions, implying that the proportion of workers in these divisions are significantly higher than the proportion that will choose to work in them if no entry barriers existed. But the ratio is larger for the higher-paid informal division which means this division has a larger proportion of involuntary workers compared to the lower-paid division. Overall, the results indicate that the formal sector should have 95% more workers than it has; the higher-paid and lower-paid informal divisions 63% and 42% less workers, respectively, if labour markets were perfectly competitive. That is, the formal sector should have more workers than it currently has, but both divisions of the informal sector should have less. Table A.3 (in appendix) shows a generally similar pattern if we define informality as non-payment of taxes on employment income. Therefore, our results are robust to two definitions of informal wage employment. In conclusion, we therefore reject the hypothesis of a fully competitive labour market, and conclude that barriers exist that prevent some workers from working in the division of the labour market where their potential earnings will be maximized.

Table 2.5: Distribution of individuals across divisions

	Formal		Informal-1		Informal-2	
	Value	[95% C.I]	Value	[95% C.I]	Value	[95% C.I]
π_j	0.386	[0.372, 0.400]	0.089	[0.080, 0.097]	0.525	[0.511, 0.540]
$\bar{\pi}_j$	0.754	[0.742, 0.767]	0.051	[0.044, 0.057]	0.195	[0.183, 0.206]
$\frac{\pi_j}{\bar{\pi}_j}$	0.512	[0.494, 0.530]	1.741	[1.482, 2.000]	2.698	[2.542, 2.855]

2.5.4 *Comparative advantage or last resort?*

Table 2.5 tells us some workers will maximize potential earnings in divisions of the labour market other than the one they are currently working in. This ultimately means that, potentially, each of the divisions will contain both voluntary and involuntary employment. This is, in fact, the case as we see in Table 2.6. A significant proportion of formal sector wage employment (93%) is voluntary compared to 28% of all informal wage employment –2% in Informal-1 and 26% in Informal-2. Therefore, workers for whom informal wage employment is a last resort account for 72% of all informal wage employment. This configuration of the labour market contrasts with the Fields hypothesis that the higher-paid (Informal-2) informal division comprised entirely of individuals who voluntarily choose to work in that division, and the lower-paid (Informal-1) comprise workers rationed out of the formal sector. Using the non-payment of employment income tax criterion of informality, the general result (see Table A.4 in appendix) holds even though the composition of voluntary and involuntary wage employment in the informal sector changes slightly to 21% and 79%, respectively.

Table 2.6: Distribution of workers across divisions where earnings will be maximized

	Formal		Informal-1		Informal-2	
Better-paid division	# of Workers	% Formal	# of Workers	% Informal	# of Workers	% Informal
Formal	1606	92.9	293	10.56	1480	53.82
Informal-1	18	1.04	52	1.89	158	5.75
Informal-2	105	6.07	52	1.89	715	26
N	1729	100	397	14.43	2353	85.67

This general picture, essentially, implies that informal sector is largely a sector of last resort for a large majority of informal wage workers. This notwithstanding, significant differences also exist in the distribution of employment types across the two informal divisions (see Table 2.6 above). More specifically, in the lower-paid division 74% and 13% of workers would be better off in the formal sector and higher-paid informal division, respectively. Therefore, a significant 87% of all workers in that division are rationed out of other divisions of the labour market. So essentially, only 13% of current workers in that division are voluntary. Similarly, in the higher-paid division 63% and 7% of workers would be better off in the formal sector and lower-paid informal division, respectively, leaving 30% of workers who are truly voluntary. Although the formal sector also contains some amount of involuntary employment (7% to be exact) this is significantly lower relative to the proportion in informal wage employment. If one considers the fact that the value of non-wage benefits, including social security, medical insurance and job security, in the formal sector is likely to be higher, relative to that in any of the informal divisions, the actual share of involuntary employment in the formal sector is likely to be close to zero.

2.5.5 Wage penalties and wage gains

The Fields hypothesis posits that workers voluntarily employed in the informal sector do so because they earn higher wages than they would earn in the formal sector. If this hypothesis holds, then the average earnings of a typical worker in voluntary informal wage employment should be significantly higher than that of an identical formal worker. With the estimated share of voluntary and involuntary employment in informal wage employment we use propensity score matching to estimate this wage gain by voluntary informal workers. In addition, we also estimate the informal wage penalty, defined as the wage differential between a typical worker in involuntary informal wage employment and an identical formal worker.

The first step in propensity score matching is estimating the propensity scores. Heckman et al. [1998] advise that variables that simultaneously impact potential outcomes and treatment status should be controlled for. Therefore, our propensity score specification model not only controls for human capital factors, but also household-level characteristics likely to impact the decision of a worker to choose involuntary employment. Results for the propensity score estimation is presented below in Table 2.7. The likelihood of involuntary informal wage employment is strongly and positively correlated with industry of employment. Indeed, given that a majority of informal wage employment is involuntary, and the fact that more than 70% of informal wage employment is located in five industries –mining, construction, manufacturing, commerce, transport, storage and communication– it is not surprising that industry is a strong predictor of involuntary informality. Education and age also strongly predict voluntary informal wage employment than they do involuntary. As expected, number of children and having an adult aged 60 years and above in household are also positively correlated with the likelihood of voluntary informal wage employment. Though, the results of the propensity score regression are interesting in themselves the primary aim is to estimate the propensity scores.

Table 2.7: Probit estimation of likelihood of voluntary and involuntary informal wage employment

Variable	Involuntary Informal	Voluntary Informal
Education *(ref: No school)*		
Low (1-6 years)	0.257 (6.24)***	-0.009 (-4.04)***
Medium (7-10 years)	0.006 (0.19)	-0.008 (-3.42)***
High (> 10 years)	0.049 (1.58)	-0.135 (-9.63)***
Age *(ref: 15–24)*		
25–44	-0.046 (-2.17)**	-0.009 (-3.30)***
45 $\geq$	-0.030 (-1.25)	-0.015 (-4.44)***
Other		
Location (Urban=1)	0.079 (5.11)***	-0.017 (-4.01)***
Sex (Male=1)	-0.053 (-3.15)***	-0.001 (-0.54)
Marital status (Married=1)	0.147 (9.42)***	-0.037 (-5.23)***
Household size	-0.015 (-3.14)***	
Infants in HH	0.006 (0.28)	
Children in HH	0.018 (2.19)**	
HH members aged $\geq$ 60	0.034 (1.78)*	
Industry		
Mining	0.457 (14.40)***	-0.005 (-3.07)***
Manufacturing	0.199 (6.87)***	0.007 (1.63)
Construction	0.081 (2.58)***	0.050 (3.61)***
Commerce	0.193 (8.34)***	0.045 (4.32)***
Transportaion & Storage & Communication	0.271 (10.54)***	0.032 (3.44)***
N	4480	4480

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Using the estimated propensity scores we estimate the average treatment effect of being in voluntary and involuntary informal wage employment, that is the (informal) wage gain and penalty, respectively. To improve the quality of matching and ensure that the treated group is as similar as the control group, we employ greedy matching to reduce the sample to closely matched pairs. Greedy matching is one way of forming matched pairs by randomly selecting a treated unit, and then matching it to a control unit whose propensity score is the closest. This process is repeated until matched pairs have been found for all treated units, or until all control units have been matched to a treated unit. An alternative to greedy matching is optimal matching. In this type of matching, matched pairs are chosen so as to minimize the total within-pair propensity score difference. Gu and Rosenbaum [1993] show that optimal matching performs as well as greedy matching in choosing balanced matched pairs.

The application of greedy matching reduces the sample for estimating the wage penalty from 3,029 observations (1299 involuntary informal and 1730 formal workers) to 796 matched pairs comprising 398 involuntary informal workers (treated units) and 398 formal workers (control units). On the other hand, the number of observations used in estimating the wage gain dropped from 2198 observations (468 voluntary informal and 1730 formal workers) to 214 matched pairs; 107 each of treated and control units. To ensure that the significance of estimates are not sensitive to the choice of matching technique three different types of matching were employed –nearest neighbor, caliper and kernel matching. The estimation results are presented in Table 2.8 below. The wage penalty is significant across all matching types. Estimates from nearest neighbor, radius and kernel matching show a penalty of GHC646.58, GHC511.16 and GHC472.18 respectively. This means that the monthly monetary earnings of a typical involuntary informal worker is on the average GHC472.18–GHC646.58 ($108.6–$148.7)[4] less compared to monthly earnings of an identical formal worker. Alternatively, this can be viewed as a tax on involuntary informal wage employment. The wage gain, on the other hand, is not significant. Again, these general results hold (though the wage penalty reduces) using the non-payment of employment income

[4] All local currency units converted at Feb, 2 2017 exchange rate of 1GHC to 0.23 US$.

tax criterion of informality (see Table A.5 in appendix). The implications of these results is discussed in the concluding section.

Table 2.8: Treatment effect (ATT)

Wage gaps	*NN-Matching* [a]	*Radius Matching* [b]	*Kernel Matching* [c]	*N*
Wage penalty	-646.58 ***	-511.16 ***	-472.18 ***	796
	(-4.22)	(-2.82)	(-5.13)	
Wage gain	18.34	-202.76	25.07	214
	(0.03)	(-0.54)	(0.03)	

Notes: Estimates obtained by bootstrap with 50 replications.
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$
[a] Number of neighbors used is 2.
[b] Caliper=0.001.
[c] Type of kernel is Gaussian, bandwidth (0.01)

The distribution of propensity scores before and after matching is presented in Figures 2.3 and 2.4. Both plots show very good balance of the propensity scores after matching; indicating that the treated and control units are similar in the two estimation samples. A covariate balance test was also performed to assess the similarity between the means of the observable covariates of the treated and control units. The results and their associated plots are presented in Tables 2.9 and 2.10 and Figures 2.5 and 2.6. According to Table 2.10 and Figure 2.6, there is no significant difference in the means of the covariates of voluntary informal workers (treated unit) and matched formal workers (control unit). Similarly, the means of the observable covariates of involuntary informal workers (treated unit) and their matched formal counterparts (see Table 2.9 and Figure 2.5) also show no significant biases. From these, we can conclude that the treated and control units of both estimation samples are adequately balance, and that observable bias (in the means of the covariates of the treated and control groups) has been sufficiently reduced.

Figure 2.3: Propensity score distribution before and after matching–Involuntary informal vs formal

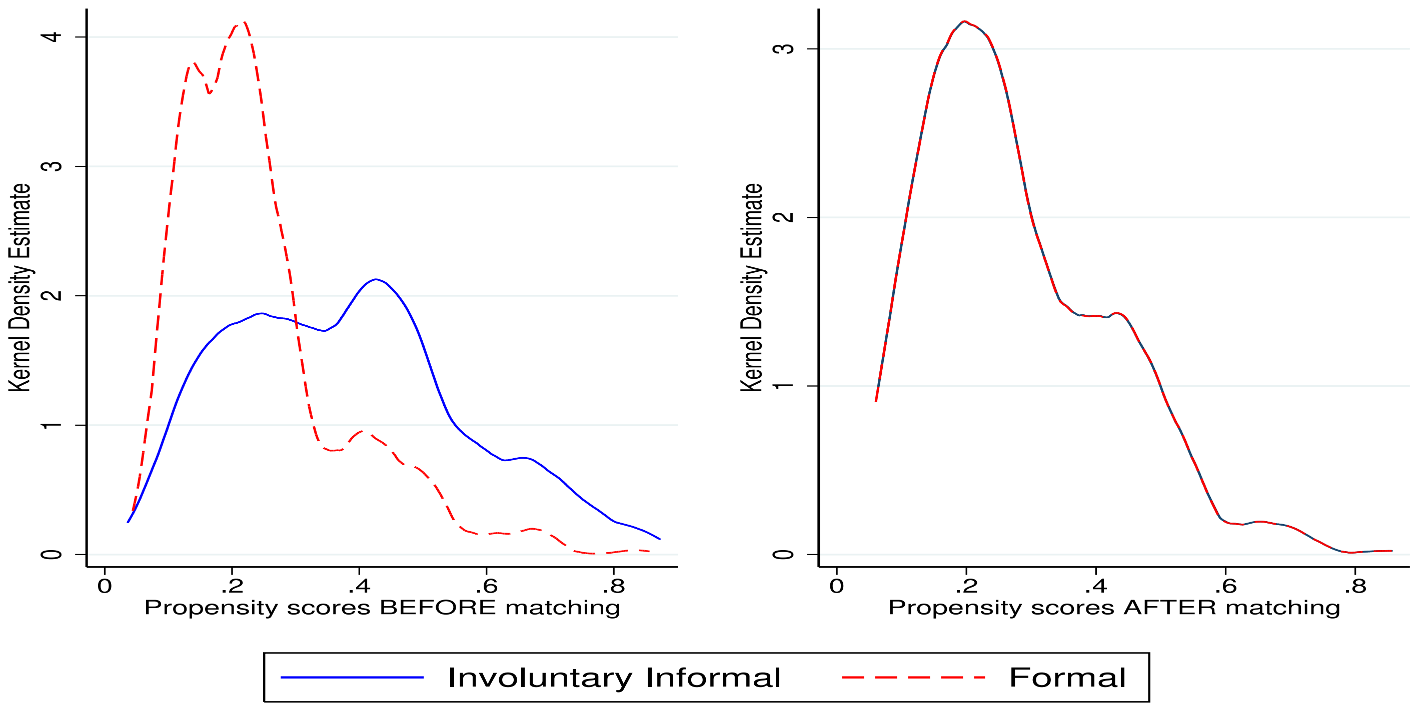


Figure 2.4: Propensity score distribution before and after matching–Voluntary informal vs formal

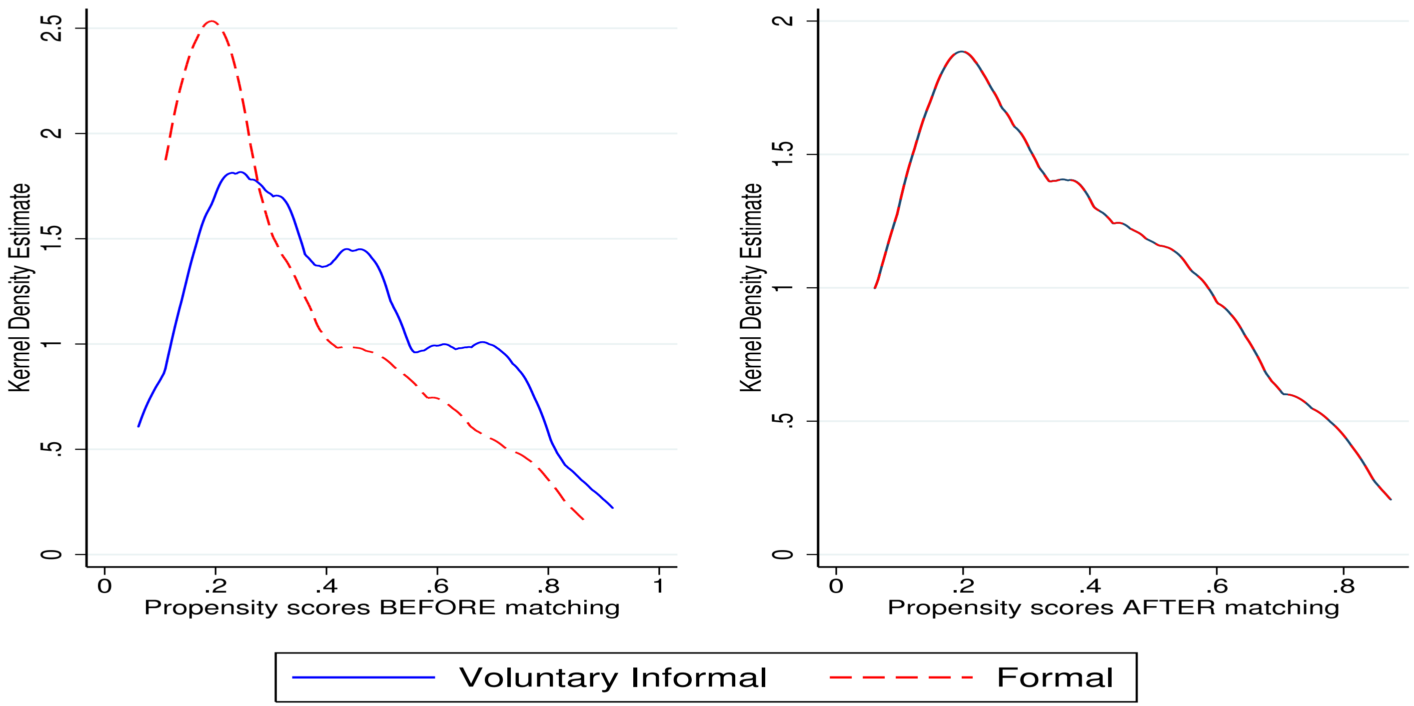


Table 2.9: T-test of covariate balance–Involuntary informal

	Mean		
Variable	Treated	Control	Bias (%)
Education *(ref: No school)*			
Low (1-6 years)	0.035	0.025	5.8 (0.81)
Medium (7-10 years)	0.287	0.270	3.9 (0.55)
High (>10 years)	0.657	0.686	-6.0 (-0.85)
Age *(ref: 15–24)*			
25–44	0.647	0.650	-0.6 (-0.09)
45 $\geq$	0.217	0.223	-1.6 (-0.22)
Industry			
Mining	0.040	0.043	-1.4 (-0.20)
Manufacturing	0.096	0.115	-6.2 (-0.87)
Construction	0.050	0.043	3.7 (0.52)
Commerce	0.156	0.140	4.6 (0.65)
Transportaion & Storage & Communication	0.126	0.121	-1.6 (0.22)
Other			
Location(Urban=1)	0.811	0.828	-4.4 (-0.63)
Sex(Male=1)	0.723	0.716	1.6 (0.23)
Marital status(Married=1)	0.637	0.643	-1.2 (-0.17)
Household size	3.433	3.479	-2.2 (-0.31)
Infants in HH	0.045	0.038	3.9 (0.55)
Children in HH	1.063	1.029	2.7 (0.37)
HH members aged $\geq$ 60	0.063	0.066	-1.1 (-0.15)
N	398	398	

Note: Test based on radius matching results
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Table 2.10: T-test of covariate balance–Voluntary informal

Variable	Mean Treated	Mean Control	Bias (%)
Education *(ref: No school)*			
Low (1-6 years)	0.093	0.095	-0.4 (-0.03)
Medium (7-10 years)	0.776	0.766	2.2 (0.16)
High (>10 years)	0.037	0.031	3.5 (0.25)
Age *(ref: 15–24)*			
25–44	0.636	0.648	-2.5 (-0.18)
45 $\geq$	0.056	0.051	2.3 (0.18)
Industry			
Mining	0.019	0.017	1.1 (0.09)
Manufacturing	0.093	0.098	-1.6 (-0.12)
Construction	0.168	0.154	3.8 (0.28)
Commerce	0.083	0.083	0.0 (0.00)
Transportaion & Storage & Communication	0.150	0.147	0.8 (0.60)
Other			
Location (Urban=1)	0.636	0.636	-0.0 (-0.00)
Sex (Male=1)	0.654	0.645	2.0 (0.14)
Marital status (Married=1)	0.224	0.235	-2.6 (-0.19)
N	107	107	

Note: Test based on radius matching results
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Figure 2.5: Test of covariate balance–Involuntary informal vs formal

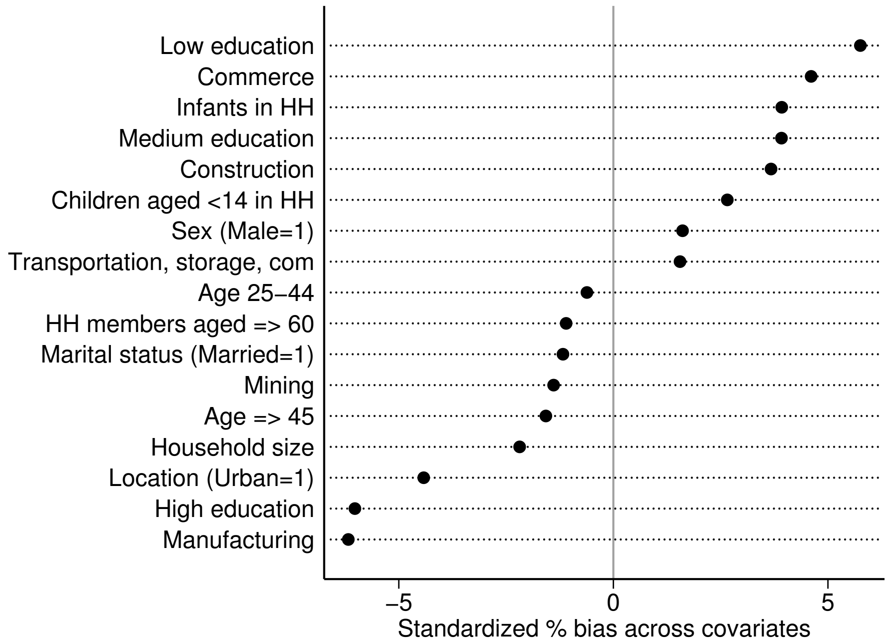
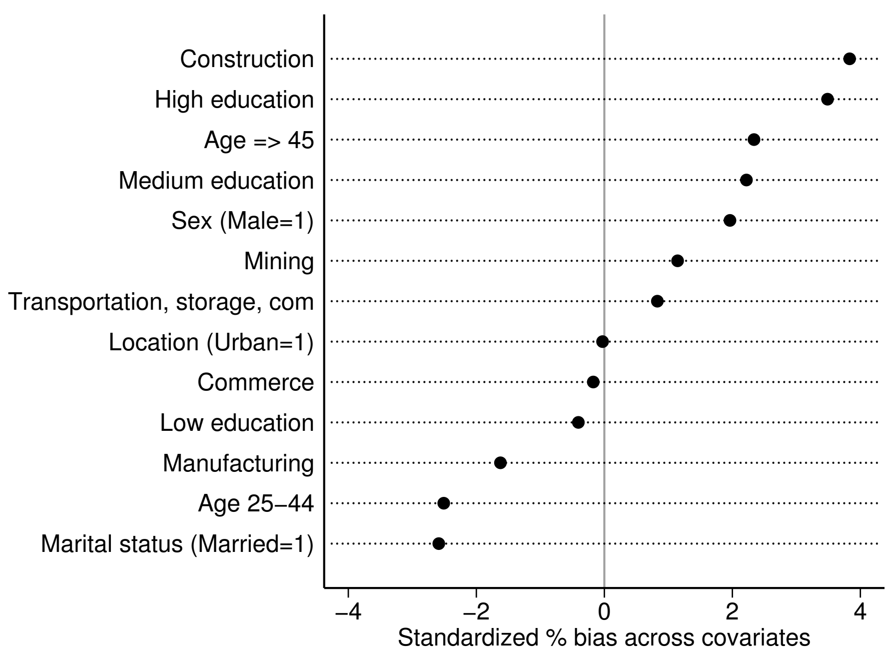


Figure 2.6: Test of covariate balance–Voluntary informal vs formal



2.6 Conclusions

This paper employs a finite mixture to model the structure of the Ghanaian labour market with regards to wage employment. The findings show that informal wage employment comprise two divisions each with a distinct wage function. This result is robust to two definitions informal wage employment –non-payment of employment income taxes and non-coverage of social security. The size of the higher-paid (Informal-1) and lower-paid (Informal-2) is, however, slightly sensitive to how informality is defined. By the social security criterion, the higher and lower-paid informal divisions make up 86% and 14% of informal wage employment, respectively. But the composition is 72% higher-paid, and 28% lower-paid if we define informality as non-payment of employment income tax. We also test whether the labour market is segmented or competitive, and we reject the hypothesis of perfect sectoral mobility, indicating that a sizeable proportion of workers in informal wage employment, 72% to 79% to be exact, will either be better off in the formal sector or the other informal division. This means, essentially, only 21% to 38% of informal wage employment is truly voluntary. This characterization of Ghanaian informal wage employment, therefore, does not fit the Fields hypothesis of a dual informal sector with an upper-tier (higher-paid) voluntary division, and a lower-tier (lower-paid) involuntary division. This results suggests several and pervasive barriers to labour mobility. These findings are similar to the results by Günther and Launov [2012] for Ivory Coast and Harati [2013] for Egypt.

Using the voluntary – involuntary split of informal wage employment, we also estimate the informal wage penalty and the wage gain. The findings indicate that workers in informal wage employment, who could maximize potential earnings in the formal sector, earn significantly less compared to identical formal workers. However, informal workers who have a competitive advantage in the sector earn as much as identical formal workers. No significant wage gain means that workers in voluntary informal wage employment are choosing the informal over formal employment for other reasons. We also found a small proportion of formal workers will be better off in informal wage employment. However, this estimate we argue may approach zero if we consider the fact that the value of non-monetary employment benefit is higher in the formal sector. The results also show, surprisingly, that approximately 8% of all workers in informal wage employment will be better off in the other division of informal wage employment. This suggests that the informal wage employment itself is either not fully competitive or each division may not have enough job opportunities to absorb these additional workers.

These findings notwithstanding one might question the robustness of the estimates with regards to the main assumption that workers are earnings-maximizers. It could be argued that workers rather maximize utility instead of earnings, and as such non–monetary employment benefits would matter. It is true non-monetary benefits matter, but there is no reason to believe that the value of these benefits will be higher in the informal sector. It is entirely reasonable to assume that non–monetary benefits in the formal sector –social security, job security, health insurance –should exceed that in the informal sector. Indeed, the distribution of non–monetary benefits in each division of the labour market is very likely to be similar to division-specific earnings distributions. As such assuming workers are utility maximizers should not significantly change the result.

In conclusion, our findings are particularly important for public policies aimed at formalizing the informal sector. First of all, policy-makers need to understand that not all informal wage employment is inferior to formal sector work, and that there is a proportion of workers in the sector who have no incentive to switch to a formal sector job. Also, the informal sector may be a cause of poverty for workers who could do better in the formal sector. For these workers, barriers to formal sector employment could just be simply due to a lack of job opportunities in the sector. In which case, government policies should be geared at expanding formal sector job opportunities. Given that involuntary informal wage employment accounts for more than 70% of all informal wage employment rapid economic growth and expansion in formal sector opportunities may be the most effective way to increase formalization and reduce poverty.

3. UPTAKE OF VOLUNTARY HEALTH INSURANCE AND ITS IMPACT ON HEALTH CARE UTILIZATION IN GHANA

3.1 Introduction

Many low and middle income countries primarily employ user fee schemes for financing healthcare. Many studies, however, have warned of the negative impact of user fee schemes on health care utilization, especially among the poorest [Preker et al., 2002; Van Doorslaer et al., 2006; WorldBank, 2004]. In the event of ill health, the required payment of user fees may deter low-income households from seeking the necessary care. Even when they do seek care many of these households may suffer financial hardships or impoverishment due to payments of large medical expenses [Peters, 2002; Xu et al., 2003]. More generally, user fees can impact the ability of households to smooth consumption over time Townsend [1994]. A key question then facing low and middle income countries is how to establish a health financing system that improves access to healthcare for all their citizens and increase financial protection. Against this background, several developing countries including Ghana, have moved to establish a social health insurance (SHI) scheme with the ultimate aim of progressively moving towards universal healthcare coverage.

Ghana established the National Health Insurance Scheme (NHIS) in 2003 to replace a user fee system that had been in operation since the 1980's. The scheme is administered by the National Health Insurance Authority (NHIA) with headquarters in Accra. Some key functions of the NHIA include licensing and regulation of district-level mutual health insurance schemes (DMHISs), accrediting providers, determining

–in consultation with DMHISs– premium levels, and general oversight of NHIS operations. The scheme is funded from three (3) main sources: a 2.5% value-added-tax on consumer goods and services, which accounts for 73.8% of total revenue, 2.5% of social security contributions from formal sector workers (20.4% of total revenue) and premiums paid by informal members (3.4% of total revenue). In addition, members have to pay a non-zero entry fee to cover the cost of insurance cards. Formal sector workers, pensioners, persons over 70 years of age, children under 18 years (if a parent is a registered member) and indigents are exempted from paying premiums. Everyone else, on the other hand, voluntarily enroll by paying a minimum income-adjusted annual premium of GHC 7.2 ($1.67)[1]. Registered members get a benefit package that covers an estimated 95% of common disease conditions in Ghana.

An efficient social health insurance scheme should have two key characteristics: mandatory enrollment, and a large number of people in the risk pool to allow for sufficient spreading of risk so as to meaningfully reduce the cost of care for those that fall ill [Varian, 1992]. Even though the NHIS law mandates enrollment for every adult, the relative size of the informal economy makes enforcement of this mandate prohibitively costly; thereby making enrollment defacto voluntary. Standard theoretical models of insurance predict two possible outcomes in the presence of voluntary insurance. First, individuals with private information about their high potential use of healthcare are more likely to self-select into insurance. Rothschild and Stiglitz [1976] show that this self-selection may lead to adverse selection where a plan that draws only high-risk individuals becomes too expensive to offer. Second, voluntary insurance may also lead to certain groups of the population, especially the poor, choosing zero or limited coverage. These poor individuals may reason that, at current prices, insurance is too expensive relative to other uses of their limited income. As a result,

[1] All local currency units converted at Feb, 2 2017 exchange rate of 1GHC to 0.23 US$.

they may choose not to buy health insurance. Indeed, given the higher concentration of poverty in the informal sector, non-zero premiums and entry fees that the NHIS charges could price out a large section of the poor, thus undermining one of the fundamental justifications for publicly provided health insurance –improving the distribution of services. In line with the theoretical predictions a large body of empirical studies have found that, in the absence of subsidies, not only do voluntary health insurance schemes fail to reach a large segment of the population, they also exclude a large majority of the poor and vulnerable [Bennett et al., 1998; De Allegri et al., 2006; Ekman, 2004; Jowett and Hsiao, 2007; Lieberman and Wagstaff, 2009; Schneider and Hanson, 2006].

Using data from the World Bank's Living Standards Measurement Survey (LSMS), this paper explores the factors that explain the uptake of voluntary health insurance by informal members of the NHIS. While a few empirical studies have examined the determinants of enrollment in the NHIS, none of these studies distinguish between voluntary and compulsory health insurance. Given that formal sector workers acquire membership of the NHIS by virtue of holding a formal sector job, the factors determining their enrollment will be correlated with factors associated with the likelihood of obtaining a formal sector job - most notably a relatively higher level of human capital. Therefore, the characteristics of a typical informal enrollee will be different from that of a formal sector worker. Failure to distinguish between the two modes of enrollment could result in misidentifying the main drivers of voluntary enrollment. Additionally, given that achieving universal coverage is akin to achieving universal coverage of the informal sector, focus should be on understanding the possible drivers of enrollment by informal members. This paper attempts to fill that gap in the literature by examining the main determinants of enrollment by informal members of the NHIS.

We employ a probit model to estimate the probability of voluntary health insurance uptake by informal members. Our findings show that, relatively older, richer and less healthy individuals are more likely to demand health insurance –supporting conclusions from previous studies that voluntary health insurance schemes may not be pro-poor. Surprisingly, we also find that smoking and alcohol consumption are negatively associated with the propensity to buy health insurance. This result may be due to the fact that individuals who smoke and/or consume alcohol are likely to be less risk-averse, than the rest of the population, and therefore less likely to buy health insurance.

The structure of the NHIS also qualifies as a natural experiment which we exploit to determine how the different channels of acquiring health insurance impacts health care utilization. Specifically, we use the method of propensity score matching (PSM) to compare the healthcare utilization of formal members (control group) with that of informal members (treated group). By comparing the frequency of annual hospital visits of informal members who buy health insurance with that of relatively observationally similar formal members, we are able to estimate the impact of voluntary health insurance on health care use. Our results show that, a typical informal member of the NHIS, on the average, visits the hospital 0.6–0.9 more days in a year compared to an observationally identical formal member.

These results are relevant to at least two aspects of the NHIS. An important goal of the scheme is to protect the poor against adverse health shocks and improve equity. But our findings show that the imposition of premiums and entry fees is excluding the poor from the scheme, which may actually worsen equity. Also, voluntary health insurance results in selection bias which may increase cost of care and reduce efficiency. Moreover, given the financing constraints faced by low income countries 'over use' of health care induced by voluntary health insurance may make these social health insurance schemes financially unsustainable.

The rest of the paper proceeds as follows. In section 3.2, we explore the theoretical underpinnings of health insurance demand. Section 3.3 surveys the existing literature on determinants of health insurance. We describe the data and variables in section 3.4. Section 3.5 addresses the empirical specification. Estimation results are presented in section 3.6, and in section 3.7 we conclude.

3.2 Theoretical Underpinnings of Health Insurance Demand

Social health insurance (SHI) schemes are motivated by perceived market failure in the private health insurance market due to asymmetric information and imperfect competition [Akerlof, 1970; Rothschild and Stiglitz, 1976; Spence, 1973], income redistribution [Cremer and Pestieau, 1996; Rochet, 1991] and paternalism [Nyman, 1999; Thaler and Sunstein, 2003]. The implication is that an efficient SHI, among others things, should reduce inefficiencies in the insurance market and increase social welfare. Toward this goal, among the solutions proposed are a combination of insurance mandates and subsidies. Mandating enrollment, the argument goes, should eliminate the inefficiency created by adverse selection by pooling low and high-risk individuals [Akerlof, 1970]. But enforcement of mandates is easier in developed economies with relatively low levels of informality. In developing countries, however, governments face major constraints in enforcing mandates because the relative size of the informal sector makes enforcement virtually impossible. This constraint is the rationale for voluntary health insurance targeted at individuals not working in the formal sector. The questions that arise then are, under what conditions would informal members seek voluntary insurance and what are the inefficiencies associated with such a scheme?

While the theoretical literature on the demand for health insurance is vast, few attempts exist to understand demand for voluntary health insurance in the informal sector [Baeza et al., 2002]. Generally, empirical studies that have examined insurance demand have employed expected utility theory (EU), due to Von Neumann and Morgenstern [1944], as the theoretical basis for explaining an individual's decision as to whether to enroll in insurance or not. According to this theory, insurance demand reflects an individual's risk aversion and the need for financial certainty [Schoemaker, 1982]. It argues that, due to the uncertainty of an illness, an individual who is risk-averse will buy insurance to protect themselves from an uncertain financial loss in the future [Arrow, 1963].

Three main implications derive from this theory. First, that a decision-maker's belief may be represented as a unique and additive probability distribution. Second, that these probabilities enter the decision-maker's preferences in a linear fashion [Kelsey and Quiggin, 1992], and third that the utility of possible outcomes are not influenced by initial endowments. The validity of these assumptions have been questioned as empirical and laboratory studies have documented systematic violations in the predictions of the theory [see Battalio et al., 1990; Kahneman and Tversky, 1979; Schoemaker, 1982]. This evidence has generated several new theories that seek to account for the weaknesses of expected utility theory. Most prominent among these are the class of reference-dependent theories such as prospect theory [Kahneman and Tversky, 1979; Tversky and Kahneman, 1992], claims theory [Mas-Colell et al., 1995; Nyman, 1999], status quo bias [Samuelson and Zeckhauser, 1988] and the endowment effect.

Prospect theory posits that people derive utility from gains and losses measured relative to a reference point. Furthermore, people weigh a loss more heavily than an equal gain, but they tend to overweight small probabilities and underweight large probabilities. Therefore, an individual will tend to be risk-seeking in losses and risk-averse in gains. That is, when faced with a risk of a loss a person will tend to exploit it rather than accept a certain loss of the same expected magnitude. This implies that individuals do not value certainty in the presence of losses, and as such demand for insurance may not necessarily be driven by a need for certainty. Instead, the demand for insurance is driven by the prospect of a gain [Kahneman and Tversky, 1979; Tversky and Kahneman, 1992]. Applied to the demand for health insurance, given a premium level, an individual first assesses his current state of health relative to a perceived future state. Then she buys health insurance if she expects to pay more for any deviation from their current state of health. However, poor individuals may choose not to buy health insurance because they may underweight the expected loss in the event of an illness [Case and Deaton, 2003].

Alternatively, contingent claims theory posits that becoming ill fundamentally impacts people's preference. Therefore, individuals may choose to buy health insurance now in order to transfer income into the ill state where the marginal utility of income is greater [Nyman, 1999]. The decision to take up health insurance is, thus, driven by expected income in the ill state, and not necessarily a need for certainty. The status quo bias and the endowment effect hypothesis, on the other hand, argue that individuals may prefer the status quo relative to something new and unknown, especially if the alternative is more complicated [Kahneman et al., 1991]. In the health insurance context, poor and low-income households may not buy health insurance because they may not understand the benefits of insurance or the concept itself might be new to them. In other words, the decision not to enroll in health insurance may be primarily due to informational problems.

Levels of risk aversion also changes with relative poverty. The poverty literature indicates that household's level of risk aversion increases as they move closer to poverty because any further decline in income could push them into –or further into– poverty [Wagstaff, 2000; WorldBank, 2000]. As such, given the relative cost of insurance, poor households may decide to choose current over future consumption and remain uninsured. Demand for health insurance will, therefore, depend on the cost of insurance, the relative values of current and future protection and the ability to smooth consumption over time in the event of an adverse health shock.

3.3 *Empirical Literature*

A large body of empirical studies have examined the factors affecting enrollment in voluntary health insurance schemes in developing countries. In general, two main methodologies have been used –discrete choice estimation techniques and contingent evaluation methods. Many of the studies that have employed discrete choice models have identified education level of the household head, religious affiliation, health status, ethnicity, marital status, number of children in household, and scheme-related characteristics such as benefit and premium levels and ease of registration as some of the factors significantly correlated with the likelihood of enrollment [Chankova et al., 2008; Jütting, 2004; Nketiah-Amponsah, 2009; Polonsky et al., 2009; Schneider and Diop, 2001; Ying et al., 2007]. The evidence, however, is mixed with regards to the relationship between income levels and likelihood of enrollment. While some studies find no relationship at all [Polonsky et al., 2009; Schneider and Diop, 2001], others find a negative association between income levels and the likelihood of health insurance uptake [Aryeetey et al., 2013; Gnawali et al., 2009; Jehu-Appiah et al., 2011].

The studies that have used the contingent evaluation method generally estimate a hypothetical demand curve for health insurance as a function of subsidies, premiums and other factors [Nguyen and Knowles, 2010]. A majority of these studies found that demand for health insurance is sensitive to premium levels and the generosity of the benefit package [Chen et al., 2002; Lofgren et al., 2008]. Distance to health care facilities is found to be negatively correlated with health insurance demand [De Allegri et al., 2006; Gumber and Kulkarni, 2000], but income and education levels have positive associations with health insurance demand [Asenso-Okyere et al., 1997; Dong et al., 2003; Zhang et al., 2006].

The impact of health insurance on health care use and financial protection has also been widely reviewed [see Acharya et al., 2012; Baeza et al., 2002; Preker et al., 2002]. In general, the evidence suggests that health insurance increases utilization of health care services among the insured [Aggarwal, 2010; Gnawali et al., 2009; Jowett et al., 2004; Mensah et al., 2010; Trujillo et al., 2005; Wagstaff, 2007]. However, the evidence on the impact of health insurance on out-of-pocket health expenditure is mixed. While some studies find significant reductions in out-of-pocket expenditures for the insured [Bauhoff et al., 2011; Ekman, 2004; King et al., 2007; Miller et al., 2009; Preker et al., 2002; Wagstaff, 2010], others find mixed [Aggarwal, 2010; Axelson et al., 2009] and no significant impact [Lei and Lin, 2009; Lieberman and Wagstaff, 2009].

In the case of the NHIS, many of the studies that have analyzed determinants of enrollments have found that economic status is the most significant predictor of the likelihood of enrollment [Aryeetey et al., 2013; Dixon et al., 2011, 2014; Jehu-Appiah et al., 2011; Singh et al., 2015]. In addition, levels of education, age, gender and geographical location are also significantly associated with the demand for health insurance. Although current health status is expected to impact the likelihood of voluntary health insurance uptake, all the studies report no significant correlation between the two variables. This may be due to the fact that the proxy variable used for current health status –self reported health– may be measured with error, and therefore does not serve as a valid measure of health status.

Membership in voluntary health insurance, according to the insurance literature, may lead to overuse of health care due to adverse selection and moral hazard [Cutler and Zeckhauser, 2000; Manning et al., 1987]. This is because high-risk individuals are more likely to enroll in voluntary health insurance, and also more likely to seek care. In developing countries, in addition to potential adverse selection and moral hazard, it is also possible that voluntary health insurance could induce additional health care use driven by loss aversion. The imposition of annual renewal fees and premiums coupled with the unpredictability of incomes in the informal sector implies that individuals who buy voluntary health insurance face a potential risk of losing insurance upon expiration of current coverage. This expectation of future possible loss of health insurance may drive increases in unnecessary care [Kahneman and Tversky, 1979; Kőszegi and Rabin, 2006]. If this effect is large and significant, then together with any selection and moral hazard effect, it could substantially increase cost of care making the object of universal coverage more difficult to achieve.

Even though a large number of studies have examined the impact of health insurance on the insured this study is the first to attempt an estimate of healthcare over use induced by voluntary health insurance. We employ the method of propensity score matching to compare the frequency of hospital visits of formal and informal members of the NHIS. Because formal sector workers face compulsory health insurance and do not pay annual premiums, then, compared to informal members, their health care utilization rates should not suffer from selection and loss aversion effects. Using propensity score matching ensures that health care use of a formal member of the NHIS is matched with that of an observationally identical informal member, thereby eliminating any potential selection bias in voluntary health insurance uptake. This allows us to attribute any difference in observed utilization rate to the effect of loss aversion.

3.4 Data

The dataset used in this study is the first wave of the Ghana Socioeconomic Panel Survey (GSPS) 2009. The GSPS is conducted by the Economic Growth Center (EGC) at Yale University and the Institute of Statistical, Social, and Economic Research (ISSER) at the University of Ghana obtained from the from the World Bank's Living Standards Measurement Survey (LSMS). In all 5,009 households were interviewed across the ten (10) regions of Ghana with regard to a broad range of socio-economic and demographic topics such as employment, education, health, labour market activities, consumption patterns and asset ownership.

Information on individual household members' health insurance enrollment status, self-reported health, smoking and alcohol consumption behavior, frequency of hospital visits, height and weight are obtained from the household health component. The demographic characteristic components contains information on age, level of education, gender, location and marital status. We proxy household wealth by constructing a linear index from asset ownership indicators using a Principal Component Analysis (PCA) as proposed by Filmer and Pritchett [2001]. PCA is a multivariate statistical technique for summarizing variability among a set of variables. Its involves an attempt to reduce a large group of interrelated variables into underlying dimensions while retaining as much as possible the variation among the variables. A total of twenty three (23) asset indicators were included in the construction of the index. These include variables on household ownership of durable goods, source of drinking water, type of toilet, source of lighting, main cooking material and type of dwelling construction material.

3.5 Empirical Strategy

3.5.1 Estimation Sample

Formal sector members of the scheme consist of registered members who work in the formal sector and contribute to social security. The formal sector is defined as economic activities, public or private, that fall under the purview of state or quasi-state regulators. Informal members, on the other hand, consist of members who pay annual premiums to register or maintain membership in the scheme. These are basically adults 18 years and above who do not work in the formal sector, and who do not qualify for exemptions.

The sample for the first regression –measuring the likelihood of voluntary health insurance uptake– consist of informal members with and without health insurance. Persons under 18 years are excluded from the sample because the NHIS categorizes these individuals as children, and they are exempted from the payment of premiums if a parent is a registered member of the scheme. Adults over 70 years, pensioners, indigents and pregnant women are also excluded from the sample because they are exempted from paying premiums. Sample means are reported in Table 3.1 below. The age distribution is similar across the insured and uninsured samples. 80% of observations in both samples are aged 25–64, 15% to 16% are 18–24 years old and about 4% 65 years and older. The distribution of the wealth index confirms our hypothesis that relatively poor households are less likely to buy voluntary health insurance. Indeed, in the insured sample only 22% of observations, compared to 36% in the uninsured sample, are in the bottom quintile –implying that poor households are overrepresented among those who do not buy voluntary health insurance. On the other hand, 77% of households in the top quintile are insured. As expected, a majority of observations in our sample are self-employed and a slightly higher percentage have health insurance. The educational distribution of the sample is equally unsurprising. Because we are basically dealing with majority informal workers, we expect a large majority of informal members to have lower levels of education. Indeed, the highest education level of more than 80% of observations in both the insured and uninsured is primary education or no formal education.

Table 3.1: Descriptive statistics (Health insurance uptake)

	Insured		Uninsured	
Variable	Mean	Std.Dev	Mean	Std.Dev
Age				
18-24	0.153	0.360	0.163	0.370
25-44	0.472	0.499	0.502	0.500
45-64	0.330	0.470	0.300	0.458
65 $\geq$	0.046	0.209	0.036	0.185
HH Wealth Index (Quintiles)				
Q1	0.219	0.414	0.360	0.480
Q2	0.237	0.425	0.238	0.426
Q3	0.179	0.383	0.169	0.375
Q4	0.188	0.391	0.131	0.337
Q5	0.768	0.382	0.102	0.302
Employment Status				
Informal employee	0.065	0.247	0.092	0.290
Self-employed	0.581	0.493	0.551	0.497
Not working	0.211	0.408	0.208	0.406
Education Level				
No school	0.559	0.497	0.654	0.476
Primary	0.312	0.463	0.269	0.443
Secondary	0.111	0.314	0.070	0.255
Post-secondary	0.018	0.135	0.008	0.087
Health Indicators				
Unhealthy (self-reported)	0.075	0.263	0.069	0.253
Smoke at least once a week	0.040	0.197	0.082	0.274

...continued

Table 3.1: Descriptive statistics, *continued ...*

	Insured		Uninsured	
Variable	Mean	Std.Dev	Mean	Std.Dev
Drink alcohol at least once a week	0.151	0.358	0.262	0.440
Underweight (BMI<18.5)	0.064	0.245	0.075	0.263
Normal weight (25 $\leq$ BMI $\leq$ 24.9)	0.554	0.497	0.638	0.481
Overweight (25 $\leq$ BMI $\leq$ 29.9)	0.271	0.444	0.217	0.412
Obese (BMI $\geq$ 30)	0.117	0.321	0.078	0.268
Region				
Coastal (Western, Central, Greater Accra, Volta)	0.224	0.430	0.409	0.492
Middle (Ashanti, Eastern, Brong Ahafo)	0.481	0.500	0.336	0.472
North (Northern, Upper East, Upper West)	0.274	0.446	0.255	0.436
Other				
Household size	4.605	2.372	4.599	2.648
Household size squared	26.828	28.232	28.162	33.478
Urban	0.412	0.492	0.300	0.458
Female	0.644	0.479	0.534	0.499
Female*65$\geq$	0.024	0.152	0.021	0.144
Married	0.722	0.488	0.679	0.467
N	2059		3702	

The sample for the propensity score estimation includes all currently registered formal and informal members of the scheme. Formal sector workers who have other private insurance are dropped because of the likelihood that their frequency of use of scheme benefits may be biased. This is because if the relative cost of private care is lower, then such individuals will use less NHIS if the two insurance schemes are substitutes. However, they will use more NHIS if private health insurance and the NHIS are complements than formal sector members will use. Premium-exempted individuals are also excluded from the sample because they neither work in the formal sector nor pay premiums to enroll. Annual hospital visits is our proxy for health care use. We report the sample means below in Table 3.2. The average annual frequency of hospital visit for a formal member of the scheme is two (2), compared to three (3) for an average informal member. This implies that, on the average, an informal member consumes about 60% more care compared to a formal member.

Table 3.2: Descriptive statistics (Healthcare utilization)

	Control		Treated	
Variable	Mean	Std.Dev	Mean	Std.Dev
Annual hospital visits	2.027	2.878	3.320	4.008
Age				
18-24	0.040	0.196	0.192	0.394
25-44	0.453	0.499	0.449	0.497
45-64	0.476	0.501	0.312	0.464
65 $\geq$	0.031	0.174	0.049	0.215
HH Wealth Index (Quintiles)				
Q1	0.062	0.242	0.178	0.382
Q2	0.076	0.265	0.190	0.392
Q3	0.071	0.258	0.197	0.398
Q4	0.240	0.428	0.189	0.392
Q5	0.484	0.501	0.185	0.388
Education Level				
No school	0.022	0.148	0.215	0.411
Primary	0.280	0.450	0.568	0.495
Secondary	0.311	0.464	0.160	0.367
Post-secondary	0.387	0.488	0.060	0.238
Region				
Coastal	0.458	0.499	0.286	0.452
Middle	0.320	0.468	0.518	0.500
North	0.222	0.417	0.196	0.397
Other				
Household size	3.627	2.054	4.511	2.283
Household size squared	17.351	19.982	25.562	26.010

. . . continued

Table 3.2: Descriptive statistics, *continued ...*

	Control		Treated	
Variable	Mean	Std.Dev	Mean	Std.Dev
Urban	0.018	0.132	0.403	0.491
Female	0.400	0.491	0.667	0.471
Married	0.716	0.452	0.657	0.475
Overweight ($25 \leq$ BMI ≤ 29.9)	0.004	0.067	0.201	0.401
Renewed at least once	0.889	0.315	0.901	0.299
At least one child in household	0.640	0.481	0.748	0.434
N	225		1890	

3.5.2 *Specification*

Health insurance decision

We model the binary decision of whether to buy health insurance or not with a probit model. Even though a person's decision to buy or not to buy health insurance is unobserved, the outcome –whether he/she actually purchases health insurance or not– is observed. According to consumer choice theory, a consumer only purchases a good if the utility of the purchase (benefit) exceeds the disutility of not making the purchase (cost). We model this difference between benefit and cost as a latent variable (y^*) such that;

$$y^* = x'\beta + \epsilon \tag{3.1}$$

(y^*) is a latent variable that represents the difference in utility derived by an individual from purchasing or not purchasing health insurance, (x') is a vector of exogenous covariates and (ϵ) is an independently and identically distributed error term $\sim N(0,1)$. The observed binary variable, y, relates to the latent variable, (y^*), such that:

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

Thus, y takes a value of 1 if an individual buys health insurance. The probability that an individual purchases health insurance, $y = 1$, can be derived as;

$$P(y = 1|x) = P(y^* > 0) = P(x'\beta + \epsilon > 0|x) = P(\epsilon > -x'\beta|x) \tag{3.3}$$

$$P(y = 1|x) = 1 - \phi(-x'\beta) \tag{3.4}$$

where ϕ is the standard normal cumulative distribution function.

The counterfactual framework and the evaluation problem

The fundamental challenge in estimating the causal impact of voluntary health insurance on health care use is to determine the level of health care use by informal members had health insurance been involuntary. In other words, we need an estimate of the counterfactual health care use of informal members. However, this potential outcome is not observed in the data; giving rise to what Holland [1986] called the "fundamental problem of causal inference". The challenge then is to use known information to impute this unobserved outcome. One possible approach could be to use the information on formal members, the control group, to estimate this missing data. But this approach is only valid under a precise condition: that the average observable characteristics of formal and informal members are statistically identical. In other words, formal and informal members are identical except the means of obtaining health insurance. We know that because informal members self-select into health insurance their mean observable characteristics will be different from that of formal members. As such, estimating the causal impact of voluntary health insurance this way will lead to biased estimates. In experimental studies, randomization works in a way that solves this selection problem. When study participants are randomly assigned to either the treated or control group then the determination of the condition to which each participant is exposed is regarded to be independent of all other variables including the outcomes [Cox, 1958; Fisher, 1925].

In observational studies, where assignment to treatment is non-random, the Nyman-Rubin-Holland counterfactual framework, also known as the Rubin Causal Model (RCM), is employed in solving the missing data problem. We formalize this framework by following the approach of Imbens [2004] as follows:

Let N units (indexed by $i = 1 \ldots N$) represent a sample drawn randomly from a large population. W_i denotes treatment status, such that $W_i = 1$ if unit i receives treatment and $W_i = 0$ if otherwise. Furthermore, each unit, i, is characterized by potential outcomes $Y_i(1)$ and $Y_i(0)$ denoting treatment and no treatment respectively. Using this notation the effect of treatment,τ_i, on unit i is given as;

$$\tau_i = Y_i(1) - Y_i(0) \tag{3.5}$$

This implies that the treatment effect, τ_i, is the observed difference between the outcomes of receiving treatment and not receiving treatment. However, for each unit, i, we only observe the outcome of receiving treatment, but not the counterfactual outcome. Hence estimating the unit treatment effect, τ_i, is impossible. But, instead we can estimate (population) average treatment effects. Our first estimand, and the most commonly studied in the literature, is the average treatment effect (ATE):

$$\tau^{\text{ATE}} = E[Y(1) - Y(0)] \tag{3.6}$$

This measures the treatment effect for a randomly selected person from the population. In a situation where treatment is only available to a segment of the population, the ATE may not be appropriate as it averages across outcomes for units that might never be subject to treatment. As such, ATE is more relevant when we are interested in how the effect of treatment applies to the general population. Given that voluntary health insurance is only targeted at individuals outside the formal sector the average treatment for the treated (ATT), which considers the effect of treatment on a randomly drawn individual from the treated group, will be more relevant. The ATT is given as:

$$\tau^{\text{ATT}} = E[Y(1) - Y(0)|W = 1] \tag{3.7}$$

The ATE is equivalent to the ATT when the distribution of covariates in the treated and control groups are identical - the exact condition achieved through randomization. But in our case because the average characteristics of a typical informal member is different from that of a formal member we expect the ATE and ATT to be different.

Methods used to estimate treatment effects depend on the kind of assumptions imposed on the treatment assignment. In social experiments, where treatment assignment is randomized, outcomes are deemed to be statistically independent of the treatment. In other words, treatment assignment is assumed to be exogenous. This follows that the distribution of observable covariates in the treated and control groups are identical. In such instances, the control group becomes a good proxy for the treated group, and thus observed outcomes of the control group can serve as a robust estimate of the counterfactual outcome $E[Y(0)]$. Under this condition, both the ATT and ATE –which are equivalent– is the difference between the mean outcomes of the treated and control groups.

In observational studies, on the other hand, self-selection into treatment leads to a breakdown of the statistical independence assumption. In such situations the 'difference-in-mean' estimator employed in the experimental case will be inconsistent. Therefore, a weaker identifying assumption is invoked to deal with the selection problem:

Assumption 1 (Conditional Independence Assumption or CIA)

$$(Y_i(0), Y_i(1)) \perp W_i | X_i \tag{3.8}$$

This assumption, also known as 'selection on observables' [Heckman and Robb, 1985] or 'unconfoundedness' [Rosenbaum and Rubin, 1983], states that conditioned on the covariates, X, assignments of units to treatment, W, is independent of potential outcomes, Y. Stated this way, the selection problem is basically treated as an omitted variable problem. As such, if all variables that influence treatment assignment and potential outcomes are controlled for then the error term should be uncorrelated with the treatment variable; ensuring that assignment to treatment is 'as good as random'. This property does not, however, imply that treatment and control units with the same ex-ante observable characteristics will have identical ex-post outcomes. Rather, this says that, but for treatment, the distribution of potential outcomes of treated units will be identical to that of control units with identical observable characteristics. An additional assumption is made regards the joint distribution of treatment and the observable covariates:

Assumption 2 (Overlap)

$$0 < P(W_i = 1|X_i) < 1 \tag{3.9}$$

The overlap or the common support condition states that units with the same observable X values have a positive probability of being in both the treatment and control group [Heckman et al., 1999]. This ensures that there is sufficient overlap of the covariate distributions of the treatment and control group, and rules out the possibility of perfect predictability of treatment assignment, W, given observable characteristics, X. When these two assumptions are satisfied, treatment assignment is said to be "strongly ignorable" [Rosenbaum and Rubin, 1983]. If one is only interested in the average treatment effect on the treated (ATT) a weaker version of these assumptions is sufficient to identify the estimator [Heckman et al., 1998; Imbens, 2004]. In this case we only need assume, $Y_i(0) \perp W_i \mid X_i$, and a weaker overlap assumption, $P(W_i = 1 \mid X_i) < 1$.

Propensity score matching

Under unconfoundedness, statistical techniques such as regression, instrumental variable and matching produce unbiased estimates of treatment effects. But matching techniques have become the method of choice because regression methods impose functional forms and distributional restrictions which may produce misleading results if these assumptions are violated. Additionally, in cases where there is a hidden selection bias, regression estimates may perform worse than matching [Rosenbaum, 2002]. In our case, because self-selection is a major problem in voluntary health insurance, it is natural to employ matching methods in our estimation of treatment effects.

The basic idea of matching involves pairing treatment and control units that are similar in terms of all relevant pre-treatment observable characteristics [Dehejia and Wahba, 2002]. When these relevant differences between any two units are controlled for in the observable covariates then the differences in the outcomes of the participants and control group can be attributed to the treatment.

For a matching technique to be successful in reducing potential bias, it has to be conditioned on the full range of observable covariates across which the treatment and control units might differ. This, however, becomes a problem in the presence of large and/or continuous covariates –the so called 'curse of dimensionality'. In their original article, Rosenbaum and Rubin [1983] suggest that when many characteristics are used in the matching process, statistical matching using propensity scores can be used to select comparison groups that are similar, on average, to participants along those characteristics. The propensity score measures the conditional probability of receiving treatment. They showed that if all information that simultaneously affect treatment assignment and potential outcomes are controlled for, then matching on propensity scores yields robust estimates of treatment effects. Matching on propensity scores, therefore, solves the dimensionality problem. Following Rosenbaum and Rubin [1983] we define the propensity score as;

$$p(x) \equiv Pr(W_i = 1|X_i = x) = E[W_i|X_i = x] \tag{3.10}$$

Where, W_i=[0,1], is the indicator of exposure to treatment for unit i, and X_i is the vector of unit pre-treatment characteristics. They show that if exposure to treatment conditioned on observable characteristics is random, then the propensity score, $p(x)$, is also random. As a result, if the propensity score, $p(x)$, is known then the average effect of treatment (ATE) can be estimated as;

$$\tau_p^{\text{ATE}} \equiv E[E(Y(1)|p(x), W_i = 1) - E(Y(0)|p(x), W_i = 0)] \tag{3.11}$$

$$\tau_p^{\text{ATE}} = E[Y(1) - Y(0)|p(x)] \tag{3.12}$$

Likewise, the average treatment on the treated (ATT) estimator will be:

$$\tau_p^{\text{ATT}} \equiv E[E(Y(1)|p(x), W_i = 1) - E(Y(0)|p(x), W_i = 1)] \tag{3.13}$$

$$\tau_p^{\text{ATT}} = E[Y(1) - Y(0)|W_i, p(x)] \tag{3.14}$$

However, it must be noted treatment effect estimates are only identified under unconfoundedness and overlap: $(Y_i(0), Y_i(1)) \perp W_i \mid p(x)$; that is, conditioned on the propensity scores treatment assignment should be independent of potential outcome.

3.5.3 *Estimation of propensity score and treatment effects*

The first step in propensity score matching is estimating the propensity scores. Discrete choice models such the probit and logit models are often preferred given that treatment assignment is typically dichotomous, and the advantage of producing predictions that are within the [0,1] bounds of probabilities. However, before estimating the propensity scores one has to choose a specification for the participation model. It is advisable to include only variables that simultaneously affect both treatment assignment and potential outcomes [Caliendo, 2006; Heckman et al., 1998]. In this regard, we include variables that affect both the likelihood of voluntary health insurance demand and the propensity to seek care such as age, household wealth, health indicators, education level and location. The results of the propensity score estimation is presented in Table 3.3 below.

Apart from the bottom wealth quintiles, marital status and the northern variables, all other variables are significant. As expected, age and education level are both negatively correlated with the likelihood of being an informal member of the NHIS. This implies that formal sector workers are likely to be relatively older and educated, on the average, compared to the rest of the population. And since voluntary health insurance is only offered to individuals outside the formal sector, older and educated individuals should be relatively less likely to enroll in voluntary health insurance. Similarly, wealthier individuals are relatively less likely to have voluntary health insurance because wealthier individuals are more likely to work in the formal sector.

Even though the result of the participation regression in Table 3.3 looks interesting its main importance is to derive the propensity scores to be used in matching the treated and the control groups. The distribution of the propensity scores is presented in Figure 3.1. The distribution of the scores before matching shows very little overlap –a concentration of scores of informal members in the right tail. This is unsurprising, given that observable characteristics of adults outside the formal sector are significantly different from that of individuals in the formal sector. Formal sector workers are likely to have higher levels of human capital compared to that of the average population. And, given that voluntary health insurance implies non-participation in the formal sector, it is unsurprising that the density plot shows little overlap. To significantly reduce selection bias and improve the robustnesses of our estimates we employ the greedy matching technique to restrict the matching sample to observations with similar propensity scores across the two groups. Utilizing this technique our matching sample reduces from 2,115 to 264 observations –132 for both control and treated groups respectively. The density plot of the propensity scores for this limited sample indicates a very good balance; implying that the treated and control groups are

Table 3.3: Propensity score model (outcome is voluntary health insurance)

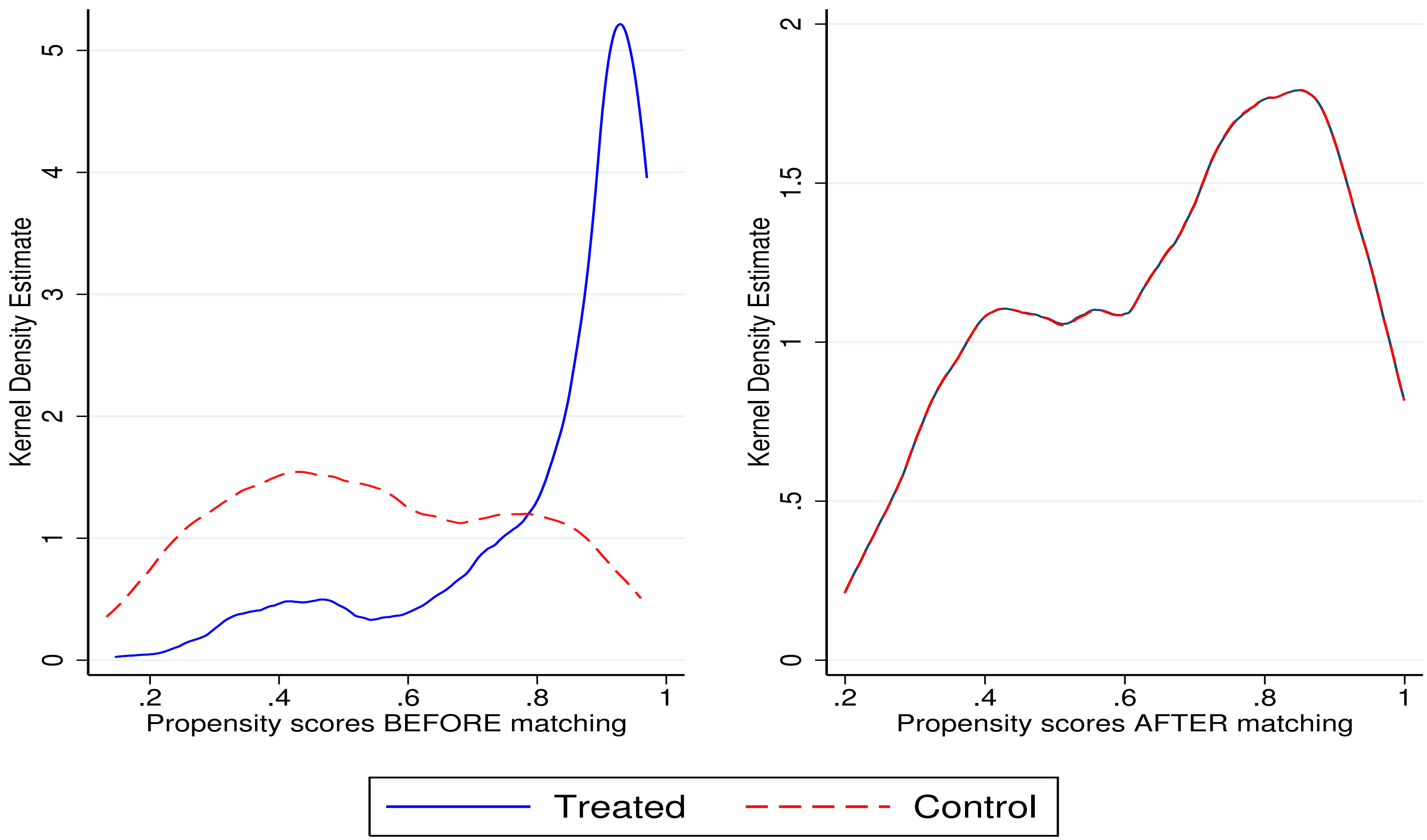
Variable	Coefficient
Constant	3.261 (4.79)***
Age *(ref: 18–24)*	
25-44	-0.029 (-2.44)**
45-64	-0.067 (-3.03)***
65 $\geq$	-0.041 (-1.19)
HH Wealth Index (Quintiles) *(ref: Q3)*	
Q1	0.011 (2.21)**
Q2	0.011 (2.13)**
Q4	-0.026 (-2.14)**
Q5	-0.040 (-2.63)***
Education Level*(ref: No school)*	
Primary	-0.024 (-2.50)**
Secondary	-0.113 (-2.94)***
Post-secondary	-0.165 (-2.94)***
Region*(ref: Coastal)*	
Middle	0.013 (2.54)**
North	-0.003 (-0.55)
Other	
Household size	0.004 (2.63)***
Urban	0.068 (5.67)***
Female	0.019 (2.99)***
Married	0.009 (1.46)
Overweight (25 $\leq$ BMI $\leq$ 29.9)	3.028 (4.72)***
Renewed at least once	0.015 (1.32)
At least one child in household	-0.008 (-1.64)
N	2115

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

similar.

In addition, a balancing test was performed to assess the covariate balance between the treated and control groups. The intent is to determine if matching successfully eliminates any significant correlation between the observable covariates and the treatment assignment.

Figure 3.1: Propensity score distribution before and after matching



The t-test for covariate balance (see Table 3.4 below) shows that all the standard mean differences in the observable covariates are not significant (at the 10% level). We,

Table 3.4: T-test of covariate balance

	Mean		
Variable	Treated	Control	Bias (%)
Age *(ref: 18–24)*			
25-44	0.481	0.471	1.9 (0.16)
45-64	0.443	0.456	-2.7 (-0.22)
65 $\geq$	0.008	0.022	-11.5 (-0.95)
HH Wealth Index (Quintiles) *(ref: Q3)*			
Q1	0.076	0.086	-3.6 (-0.3)
Q2	0.107	0.124	-5.4 (-0.44)
Q4	0.229	0.233	-1.0 (-0.08)
Q5	0.374	0.388	-2.8 (-0.23)
Education Level*(ref: No school)*			
Primary	0.404	0.404	0.0 (0.00)
Secondary	0.313	0.303	2.0 (0.16)
Post-secondary	0.260	0.258	0.3 (0.02)
Region*(ref: Coastal)*			
Middle	0.473	0.418	11.2 (0.90)
North	0.153	0.170	-4.7 (-0.38)
Other			
Household size	3.970	4.043	-3.6 (-0.29)
Urban	0.023	0.023	0.0 (-0.00)
Female	0.374	0.405	-6.4 (-0.52)
Married	0.771	0.782	-2.5 (-0.21)
Overweight ($25 \leq$ BMI ≤ 29.9)	0.00	0.005	-5.8 (-0.82)
Renewed at least one	0.916	0.923	-2.5 (-0.21)
At least one child in household	0.656	0.689	-6.8 (-0.55)
N	132	132	

Note: Test based on radius matching results
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

therefore, conclude that the treated and control groups are well balanced, and that bias from observables has been substantially and sufficiently (in a statistical sense) reduced.

To ensure that the estimated treatment effect is not sensitive to the choice of technique we tried various matching techniques; nearest neighbor, caliper and kernel matching, as well as, propensity score stratification and OLS regression (on the matched sample). The estimates from all the techniques are significant and is discussed in the next section.

3.6 Results

3.6.1 Health insurance uptake

In Table 3.5 below, we report the average marginal effects from the regression of the likelihood of voluntary health insurance uptake. This regression was run on a sample of all adults 18 and above who do not work in the formal sector. The results confirm our hypothesis that older adults are more likely to buy health insurance. Specifically, adults aged 65 and above have a 54% higher probability of buying health insurance compared to 18–24 year olds. The probability of voluntary health insurance also rises with wealth. Relative to the bottom quintile, the likelihood of buying health insurance increases as we move up the income quintile. This results supports our initial hypothesis, as well as, conclusions from previous studies [Aryeetey et al., 2013; Jehu-Appiah et al., 2011; Nketiah-Amponsah, 2009], that the NHIS may not be pro-poor. The implication is that the scheme may actually work to worsen inequality between the rich and the poor.

Level of education is also positively associated with the likelihood of buying health insurance. Again, this is unsurprising because individuals with higher education may better understand health insurance, as well as, having the capacity to access. A puzzling result, however, is the negative association between smoking cigarettes and drinking alcohol and the likelihood of buying health insurance. A possible explanation for this result may be that these individuals are less risk averse compared to the general population. This is a reasonable assumption given that smokers and people who consume alcohol are also more likely to discount the future less. As such, they will be more risk loving, relative to the general population, and, therefore, less likely to buy health insurance. Overall, the results confirm our hypothesis that the better educated, wealthier and older individuals will self-select into (voluntary) health insurance.

Table 3.5: Probit estimation of health insurance uptake

Variable	Coefficient
Constant	-1.204 (-10.04)***
Age *(ref: 18–24)*	
25-44	-0.057 (-0.95)
45-64	0.139 (2.20)*
65 $\geq$	0.542 (3.71)***
HH Wealth Index (Quintiles) *(ref:Q1)*	
Q2	0.367 (7.31)***
Q3	0.484 (7.76)***
Q4	0.626 (9.24)***
Q5	0.672 (8.33)***
Employment Status	
Informal wage employee	-0.0375 (-0.45)
Self-employed	0.119 (2.16)*
Not working	0.049 (0.79)
Education Level *(ref: No school)*	

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

. . . continued

Table 3.5: Probit estimation of health insurance uptake, *continued . . .*

Variable	Coefficient
Primary	0.196 (4.40)***
Secondary	0.334 (4.67)***
Post-secondary	0.501 (2.97)**
Body Mass Index *(ref: Underweight)*	
Normal weight ($25 \leq$ BMI ≤ 24.9)	-0.035 (-0.47)
Overweight ($25 \leq$ BMI ≤ 29.9)	0.092 (1.16)
Obese (BMI ≥ 30)	0.055 (0.60)
Other Health Indicators	
Unhealthy (self-reported)	0.171 (2.45)*
Smoke at least once a week	-0.246 (-2.98)**
Drink alcohol at least once a week	-0.260 (-5.49)***
Region *(ref: Central)*	
Middle	0.606 (14.27)***
North	0.824 (14.21)***
Other	
Household size	0.049 (2.26)*
Household size squared	0.005 (-2.92)**
Urban	-0.019 (-0.38)
Female	0.232 (5.65)***
Female*65$\geq$	-0.422 (-2.32)*
Married	0.183 (3.95)***
N	5761

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

3.6.2 Impact on health care utilization

Table 3.6 below presents the estimates of the impact of voluntary health insurance on health care use. We estimated the impacts using propensity score stratification, propensity score matching: radius, nearest neighbor and kernel and OLS. In propensity score stratification, the estimated propensity scores are used to stratify the sample into subclasses with similar propensity scores. The treatment effect is then estimated as a weighted sum of the differences of sample means across strata. Rosenbaum and Rubin [1983] show that perfect stratification based on the propensity scores will generate strata where the average within strata treatment effects is an unbiased estimate of the true treatment effect. The advantage of this approach is that the entire sample is used in estimating the treatment effect. According to Cochran [1968], often five (5) strata based on the propensity score removes over 90% of the bias in the observable covariates. To determine the optimal number of strata we use the stata program *pscore* [Becker and Ichino, 2002] which simultaneously estimates the propensity score and tests for the optimal number of strata. Based on the results, nine (9) strata or subclasses were used. The result from this techniques indicates an average treatment effect of 0.7; significant at the 5% level. This implies that, on the average, voluntary health insurance increases annual hospital visits of informal members by approximately 0.7 days.

Table 3.6: Treatment effect (ATT)

	Stratification Method [a]	*NN* [b]	*Radius* [c]	*Kernel* [d]	OLS
ATT	0.670**	0.650**	0.878**	0.645*	0.711**
	(1.95)	(1.95)	(1.97)	(1.54)	(2.80)
N	2115	264	264	264	264

Notes: Estimates obtained by bootstrap with 50 replications.
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$
[a] Number of strata used is 9
[b] Number of neighbors used is 2.
[c] Caliper=0.01.
[d] Type of kernel is Gaussian, bandwidth (0.01)

To reduce bias in our estimate we employ greedy matching to limit the sample to those observation with good matches across the control and treated group. This ensures that our matched samples are adequately balanced on the propensity scores. Using this approach, all but 264 observations were dropped –132 from both treated and control groups. The large number of observations dropped implies that any estimated treatment effect may apply only to this limited sample. Results from the matching estimates indicate significant impact of voluntary health insurance on health care use. The second column of Table 3.6 presents the result of nearest neighbor matching with replacement. This matching procedure involves matching a treated individual with an individual from the control group in terms of the closest propensity scores. The result shows that voluntary health insurance significantly increases annual hospital visits by an average of 0.7 days.

A major disadvantage of nearest neighbor matching is the risk of bad matches in cases where the closest neighbor is far away [Caliendo, 2006]. Radius or kernel-based matching solves this problem. Radius matching involves matching an individual from the treated group with observations in the control group that lie within a radius, and are closest in terms of the propensity score. Kernel matching, on the other hand, uses all observations in the control group to construct the counterfactual outcome such that it places more weight on matches that are nearer by propensity scores and less weight on distant matches [Guo and Fraser, 2014]. This has the advantage of increasing efficiency due to more observations being used in estimating the counterfactual outcome. Results from radius and kernel matching (see second and third columns of Table 3.6) indicate an approximate average impact of 0.9 and 0.6 days, respectively. This implies voluntary health insurance, on the average, increases annual hospital visits by approximately 0.6-0.9 days.

Finally, we check if OLS estimates on the matched sample is consistent with estimates from the propensity score matching and stratification approaches. The result illustrates that voluntary health insurance increases annual hospital visits by 0.7 days; not significantly different from the average estimate from PSM and stratification methods. Given the significance and consistency of the estimates across the different approaches, we can say with confident that buyers of voluntary health insurance consume more health care (in the form of a higher number of hospital visits).

3.6.3 Sensitivity analysis

Under unconfoundedness matching estimators produce unbiased estimates of treatment effects. However, if there are unobserved variables that simultaneously affect treatment assignment and potential outcomes a hidden bias might arise [Rosenbaum, 2002]. In such situations, matching estimators are no longer robust. We, therefore, employ the approach proposed by Rosenbaum [2002] and implemented by DiPrete and Gangl [2004] to test whether there are unobserved variables that simultaneously impact the likelihood of buying voluntary health insurance and the frequency of hospital visits.

The Rosenbaum method is predicated on the assumption that the effect of a potential hidden bias on the treatment estimates can be measured relative to a sensitivity parameter (Γ). A study is said have no hidden bias if Γ=1, and as the level of Γ increases, so does the level of the bias. Because the true value of Γ is unknown, the analysis proceeds by assessing how large Γ has to be before the significance of our estimated treatment effects change. The maximum Γ at which this hypothesis holds is called the "Rosenbaum bounds" and it provides bounds on the estimate.

Result of the sensitivity analysis is presented in Table 3.7 below. The estimates reveal that our treatment estimate becomes sensitive to hidden bias at $\Gamma > 1.5$. This implies that two similar individuals with the same sets of observable covariates could differ in their odds of receiving treatment by as much as a factor of 1.5 and greater because of different values on an unobserved covariate. In other words, conditioned on the covariates two identical individuals could differ in the odds of buying voluntary insurance occasioned by an unobserved covariate such that one could be 1.5 times more likely to buy voluntary health insurance. This is unlikely because an informal member with observable characteristics identical to that of a formal sector worker has as much a chance of working in the formal sector and subsequently obtaining health insurance. We, therefore, conclude, given the relative higher value or Γ, that the estimated treatment effect is not significantly sensitive to hidden bias.

Table 3.7: Sensitivity analysis

Γ	*p value*
1	0.0001
1.5	0.0871
2	0.5703
2.5	0.9145
3	0.997
N	264

Notes: Deltas calculated based off kernel matching estimates

3.7 Conclusion

Taken together, the results support our hypothesis that the poor may be excluded from health insurance due to their inability to afford premiums. Rather, we found that older people, individuals from richer households and people with higher education are more likely to buy health insurance. This association is consistent with much of the findings from previous studies [Aryeetey et al., 2013; Jehu-Appiah et al., 2011; Nketiah-Amponsah, 2009]. Consequently, to improve equity in access to health care, adjustments will need to be made to extend health insurance to the poor.

Our findings also show that voluntary health insurance is causing over-use of health care services. Though the average estimated impact of 0.7 days may seem marginal, relative to the average annual hospital visits of the control group (2.027), voluntary health insurance uptakes increases annual hospital visits by 35%. One potential explanation for this increase is loss aversion as applied in reference-dependent theories [Kahneman and Tversky, 1979; Kőszegi and Rabin, 2006]. Given that individuals who buy voluntary health insurance pay annual premiums, it is reasonable to expect that if an individual believes they may be unable to afford insurance upon expiration of current coverage they will be more likely to use health care before current premiums expire. Formal sector workers, on the other hand, because they do not pay premiums are able to smooth out their health care consumption. If this assumption holds then an informal member, on the average, should use more health care relative to an identical formal member.

While we are confident in our findings our paper has a limitation. We do not claim that the cost of health care by an average informal member exceeds that of an identical formal worker. Since we have no data about the cost and type of treatment sought by individuals we are unable to determine if increased use of health care actually translates into higher cost of care. However, given that claims payment is based on quantity, and not quality, of care offered it is likely that increased health care use will increase health care cost. Despite this limitation, this study makes an important contribution to the literature by showing that the mode of acquiring health insurance may matter to health care use.

4. THE GENDER WAGE GAP IN GHANA: EMPIRICAL EVIDENCE FROM THE FORMAL AND INFORMAL SECTORS

4.1 Introduction

The existence of a gender wage is a persistent reality across the world. Many studies have been conducted to account for the magnitude, persistence and evolution of this gap and its impact on poverty alleviation. Besides the equity concerns, analyzing and adequately measuring the gender wage gap is critical from a poverty reduction perspective. In order to formulate appropriate policies aimed at bridging the gap it is useful to explore what may be causing it. While the literature on the gender wage gap in advanced western economies is large and well developed that of Africa, in particular, is relatively modest. Weichselbaumer and Winter-Ebmer [2005] document that the largest share of the literature on developing countries examines Asia. Latin America has the second largest and Africa has the smallest –3% of all global studies since 1990.

A distinguishing feature of labour markets in Africa is the existence of a large 'unproductive' informal sector alongside a relatively smaller productive formal sector. It is estimated that the informal sector accounts for 72% of non-agricultural employment in Africa [ILO, 2008]. Two main hypotheses explain the segmentation that characterizes labour market in Africa. According to the traditional view [Dickens and Lang, 1985; Fields, 1975], segmentation arises because wages in the formal sector are set above market-clearing due to minimum wage laws, efficiency-wage explanations or higher unionization. A consequence of this is that workers in the informal sector are rationed out of the formal sector, resulting in the average informal worker earning relatively less than an observationally identical formal sector worker. However, results from Chapter 2 and evidence from some other studies show that a significant proportion of informal sector jobs actually reflect voluntary choices, that some workers are actually choosing jobs in the sector given their skill endowment, preferences, and the earning potential in the sector [Maloney, 2004; Saavedra and Chong, 1999; Yamada, 1996]. This explanation has a major implication on the dynamics of any gender wage gap that may exist within the two sectors. For instance, if labour markets are competitive then it is more likely that the gender wage gaps within sectors should not be significantly different after accounting for compensating differentials. This is because in a competitive market discriminated women in any sector would move to the other sector with no or less discrimination, and this will continue until discrimination is equalized. On the other hand, a significantly different and persistent wage gap within sectors may suggest the existence of rigidities and the need for policy action to support workers that are disadvantaged. A persistent wage gap will also mean that labour markets are not competitive and government intervention may increase social welfare. In this light, accurate measure of the gender wage gap within sectors will represent an important aspect of the analysis of labour markets in the developing world.

A majority of the African literature that have explored the gender wage gap have either employed the canonical Oaxaca [1973] decomposition or a variant of it to decompose the gender wage differential into a portion explained by human capital factors such as education and experience, and a portion due to differences in coefficients or discrimination. However, an important limitation of this decomposition technique is that it focuses on average effects which may lead to incomplete assessment of the wage gap if the gap varies significantly along the wage distribution. This in particular may be the case in the informal sector due to its heterogeneity, which means the wage gap may have different magnitudes along the wage distribution function.

In this regard, the contribution of this paper is twofold. First, we look at the differences and similarities in the gender wage differentials in the formal and informal sectors and how they vary along their respective conditional earning distributions using the quantile regression technique proposed by Machado and Mata [2005] with data from the Ghana Living Standards Survey (GLSS) VI. This technique permits the decomposition of the wage gap into effects due to characteristics and coefficients at different quantiles of the wage distribution function. We also attempt to control for the non-random participation of women in employment.

To the best of our knowledge this paper is the first empirical study that attempts a comparative analysis of the gender pay gap in the formal and informal sectors of the Ghanaian economy using quantile regression with correction for sample selection.

The results indicate an increasing wage gap along the conditional wage distribution in the informal sector. Specifically, the result shows that, on the average, approximately 16%–18% of the (conditional) wage gap is explained by differences in characteristics with the remaining of 82%–84% attributable to differences in coefficients. After correcting for positive selection into informal wage employment by women, we find that the wage gap widens; increasing, on the average, by 7%–12% along the conditional wage distribution. In the formal sector, however, not only are the magnitudes of the pay differences smaller, relative to the informal sector, there was no significant wage gap at the lower quantiles. After correcting for selection, we found that there is a positive wage gap at the lower to median quantiles but this disappears at the upper quantiles. These results collectively indicate that women are relatively better off in the formal sector. We discuss the policy implications of these results.

The rest of the paper is organized as follows. The next section presents a brief literature review. Section 4.3 describes the data, variable selection and identification of formal and informal employment. Section 4.4 presents the econometric methodology and sample selection. Empirical results are discussed in section 4.5, and section 4.6 concludes.

4.2 Literature Review

There are two main theories of discrimination. The first is the taste-based discrimination theory that originates from Becker [1957]. This model attributes discrimination to the aversion of employers, coworkers or customers towards members of a certain group. In this model, employers may have a 'taste for discrimination'. This implies there is a disutility to employing workers from a certain group. As a consequence, these workers may have to compensate employers by being more productive at the going wage or accept a lower wage for identical productivity. However, if labour markets are competitive, in general equilibrium discrimination can only be sustained at a positive cost to the employer. This is because each worker must be paid a wage equal to their marginal product in a competitive market. Therefore, non-discriminating firms will out-compete discriminating firms by offering a slightly higher wage for the services of discriminated workers. Because the discriminated workers are already being paid below their productivity employers should still make a profit. The discriminatory employer will end up losing all discriminated workers to the non-discriminatory employer, ending up with a more expensive labor and less profit. In the end, the business of the discriminatory employer becomes less profitable and goes out of business, and the wage difference due to discrimination eventually goes away Schippers [1987].

Coworkers and customers can also be sources of discrimination in this model. If workers who belong to a majority group are prejudiced against workers from a minority group and they demand a premium, at the expense of minority workers, from employers in order to work with minority workers this becomes a source of discrimination. Clearly, this situation cannot arise under perfect competition where perfect mobility of workers will ensure that this discrimination is eliminated. Customer discrimination, on the other hand, arises where customers discriminate against members of certain group, and as such get lower utility if they purchase services from firms that employ workers from this group. As a result, workers from this group may have to accept lower wages in order to work in jobs with customer contact. This kind of discrimination may still exist even if markets are competitive because there is no mechanism for consumers to compete the discrimination away. In general equilibrium, prejudiced customers may bear the cost of discrimination by paying higher prices for goods and service.

The second prominent model of discrimination is Arrow [1973] and Phelps [1972] 'statistical discrimination' model. This model seeks explanation for discrimination in incomplete information. The simplest explanation involves the idea that measuring worker productivity is a difficult task. Therefore, employers tend to base their hiring decisions on past experiences. Given that employers will tend to have better knowledge of the productivity of a majority group than a minority group, they (employers) tend to offer lower wages to minority groups. However, in the long run discrimination should go away because employers will be able to observe the productivity of discriminated workers and offer wages that matches that. Another example of statistical discrimination arises from employers' belief from past experiences that because of child bearing young female workers have less labour market attachment than men. This, therefore, results in less human capital formation in women relative to men leading to lower wages for women on the average. Statistical discrimination models, however, fail to explain the persistent wage gap we observe among groups with identical productive capabilities. This is because if individual performance is independent of group membership then repeated observation of workers' productivity should allow employers to arrive at its true value and sooner than later statistical discrimination will be remedied [Arrow, 1998; Cain, 1986]. However, this will not be the case if the negative prior beliefs of employers become self-fulfilling and act as an incentive for workers not to acquire education [Coate and Loury, 1993; Lundberg and Startz, 1983].

A majority of the empirical research that has attempted to estimate the gender wage gap has done so with reference to the discrimination theory. Therefore, the usual approach in accounting for the wage difference has been to decompose the wage gap into a portion explained by productivity differences and an unexplained portion due to discrimination [Blinder, 1973; Oaxaca, 1973]. A plurality of the empirical work has been based on data from developed countries, mostly the OECD countries. In the developing world, Asia has the largest share of the literature [Ashraf and Ashraf, 1993; Fan and Lui, 2003; Horton, 2002] followed by Latin America [Montenegro, 2001; Psacharopoulos and Tzannatos, 1991], and then Africa. The existing literature on Africa suggests a broad consensus regards the existence of the wage gap. For instance a study by Glick and Sahn [1997] indicates that, in Guinea, differences in endowments explain 45% and 25% of the gender wage gap in self-employment and public-sector employment, respectively. However, in the private sector, women actually earn more than men. In Tanzania, Armitage and Sabot [1991] find a wage gap in public sector, but observe no significant wage gap in Kenya's labour market. Likewise, Appleton et al. [1999] analyzed the wage gap in the public sector of three African countries – Ethiopia, Cote D'Ivoire, and Uganda– and concluded that the wage gap is narrower in these countries than it might be otherwise because women are overrepresented in the public sector. Agesa [1999] examines Kenyan data on formal sector workers and the self-employed, and concluded that there is a significant gender wage gap in favor of men. Hinks [2002], Nordman and Wolff [2009] are other studies on Africa that found a significant wage bias in favor of men.

The literature on Ghana reflects the scantiness of the African literature. Few empirical studies have examined the gender wage gap in the Ghanaian labour market, and all but Addai [2011] have focused on the formal sector. Verner [1999] used matched employer-employee data to estimate plant level production and wage functions which was later used to decompose the wage differential. The paper found that females are paid significantly less than males. In a related study, Glewwe [1996] found no wage discrimination against women in the private sector. However, he found that women appear to be better off than men in the public sector. More recently, Addai [2011], the only study to examine the gender pay gap in the informal sector, reported a significant wage gap in the informal sector in favor of men despite the average female being better skilled than an identical male. Without downplaying the significance of these results, we argue that a re-examination of the gender wage gap in Ghana in order. This is because previous studies have employed OLS or the Oaxaca-Blinder (OB) (1973) decomposition technique which allows for estimating wage gap only at the mean. This overlooks the fact that the gap varies along the entire wage distribution; which is what we find in our results. Given that the Ghanaian economy is about 80% informal [Osei-Boateng and Ampratwum, 2011] emphasis should be placed on examining that sector's wage gap in order to help formulate appropriate labour market policies aimed at helping the working poor. Even though the study by Addai [2011] examined the pay gap in the informal sector, it only looked at informal employment in the city of Kumasi, and as such does not paint a complete picture of the pay gap in the entire labour market. Our attempt to compare the gender pay gap in the formal and informal sector is to fill in the missing gaps in the previous literature as outlined above and also help throw light on the dynamics between the two sectors of the Ghanaian labour market.

4.3 Data

This paper also uses data from the sixth round of the Ghana Living Standards Survey (GLSS) conducted in 2012/2013 [1]. The GLSS is a nationally representative survey designed to generate information on the living conditions in Ghana. Individual members of households are interviewed over a broad range of socioeconomic topics including education, health, labour market activities, access to financial services and asset ownership. The GLSS V1 was selected because it is the most recent and comprehensive survey on the labor force of Ghana. It also comes with a larger sample size relative to previous rounds.

4.3.1 Determination of formal and informal employment

The formal sector is defined as economic activities that fall under the purview of state regulation. These activities can either be public or private and employment in them is termed formal employment. In this paper, formal employees will comprise employees who pay taxes on their wages and who benefit from or contribute to social security or pension. Defining informality, on the other hand, is not straightforward. In the labour market literature, however, two main definitions have been employed –the productivity and legalistic definitions. In this study, we employ the legalistic definition of informality and focus on informal wage employment because the GLSS VI survey contains precise questions that enables for maximum sample extraction from the data based on this criterion. The legalistic definition defines informal wage workers as those who do not pay taxes on employment income, are without a written employment contract or have no social security coverage [Hussmanns, 2004]. We also define wage to include the monetary value of wage payments in goods and services.

[1] See chapter 2 for a more detailed description of the GLSS VI.

4.3.2 Sample selection

Our formal sector sample size is restricted to workers aged 15–64 who pay taxes on their employment income or contribute to or benefit from social security or pension (4,373 observations). The informal sample, on the other hand, comprise workers aged 15–64 who do not pay taxes on their employment income and also do not contribute or benefit from social security or pension (13,204 observations). Workers with non-observable wages are excluded. Self-employed workers are also dropped because they are not the focus of our study. Additionally, students and casual workers are dropped from the sample leaving only workers who self-reported as paid employees with non-zero observable wages. After these exclusions, the formal sample comprised 1,118(68%) men and 518(32%) women) while the informal sample had a total of 2,344 working adults –comprising 1,613(69%) men and 731(31%) women.

4.3.3 Distribution of wages

The wage measure used in this paper is the natural logarithm of hourly wage. We first estimated total monthly wage and then adjusted it according to the reported monthly hours worked. Total monthly wage is the sum of monetary and the value of non-monetary payments. Tables 4.1 and 4.2 below present means and standard deviations of the variables in the study for the formal and informal sectors, respectively. Mean of log hourly wage for men in the formal sector is 2.656, and that of women with selection correction is 2.727. In the informal sector, on the other hand, the mean of log hourly wage for men 1.861 and that of women with selection correction is 1.130. On the average, women make 6.8% more than men in the formal sector, while men make twice as much as women in the informal sector. This implies the average woman in the formal sector earns a higher wage relative to the average man in the sector. In the informal sector, on other hand, the average man out-earns the average woman. These estimates hide important details on the variation of the wage gap along the entire wage distribution function.

Table 4.1: Descriptive statistics (Formal Sector)

	Men		Women		Women (w/selection)	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
Log of hourly wage	2.656	0.772	2.550	0.753	2.727	0.669
Education						
No school	0.017	0.129	0.006	0.076	0.002	0.049
Primary	0.050	0.216	0.021	0.144	0.017	0.128
Secondary	0.386	0.487	0.264	0.441	0.230	0.421
Post-secondary	0.552	0.498	0.687	0.464	0.742	0.438
Industry (major sectors)						
Education, Health & Social Work	0.381	0.486	0.649	0.478	0.661	0.474
Manufacturing	0.070	0.255	0.041	0.197	0.024	0.152
Mining & Construction	0.072	0.258	0.010	0.108	0.009	0.097
Public Admin & Finance	0.159	0.366	0.135	0.342	0.140	0.347
Electricity & Utilities	0.008	0.089	0.008	0.088	0.007	0.084
Firm Size						
< 10	0.275	0.447	0.221	0.415	0.219	0.414
> 10	0.364	0.481	0.412	0.493	0.420	0.494
Other						
Location (Urban=1)	0.631	0.483	0.792	0.407	0.815	0.389
Age	32.80	10.62	37.91	11.26	38.20	10.62
Age squared	1189	799.5	1564	925.8	1572	876.2
Marital status (Married=1)	0.568	0.496	0.552	0.498	0.602	0.490
Tenure	6.862	6.798	10.56	10.30	11.09	10.35
Children	0.380	0.190	0.533	0.499	0.642	0.480
N	1118		518		422	

Table 4.2: Descriptive statistics (Informal Sector)

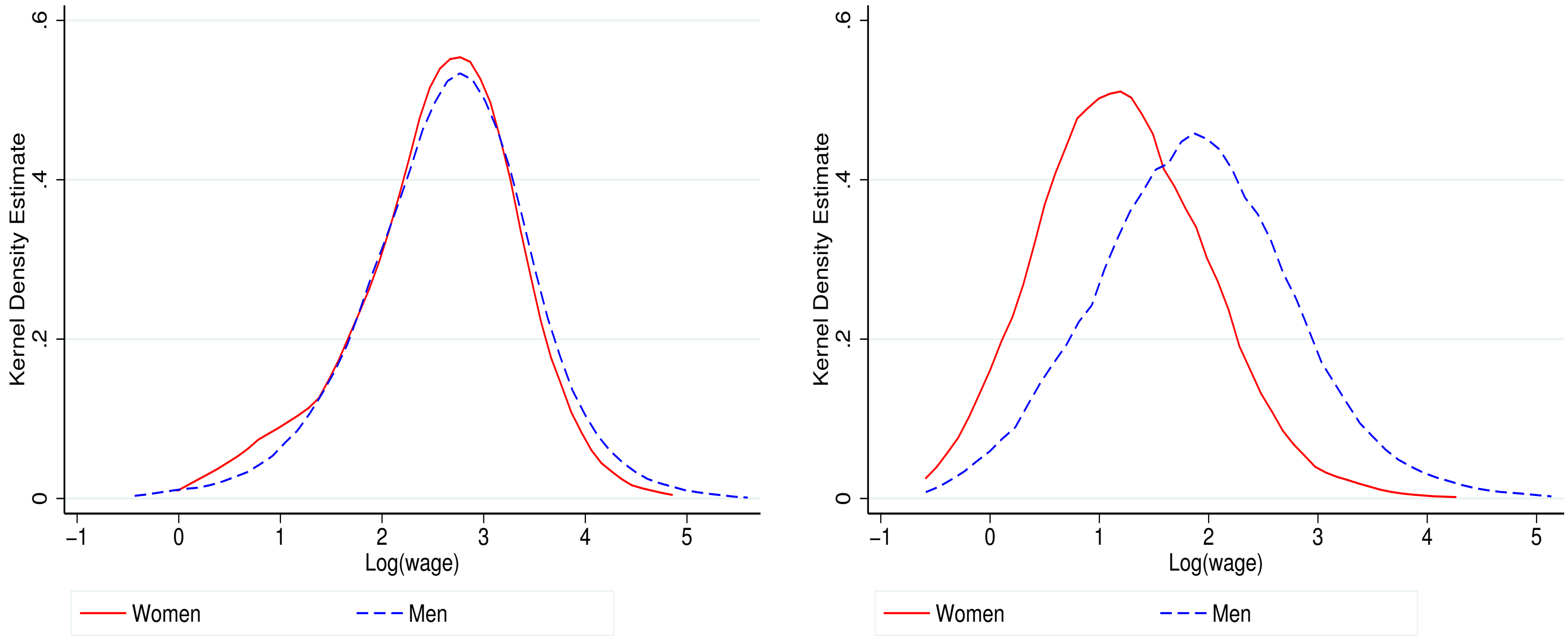
	Men		Women		Women (w/selection)	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
Log of hourly wage	1.861	0.863	0.122	0.720	1.130	0.708
Education						
No school	0.092	0.290	0.097	0.296	0.330	0.339
Primary	0.138	0.345	0.157	0.364	0.234	0.424
Secondary	0.605	0.489	0.488	0.500	0.446	0.498
Post-secondary	0.074	0.262	0.094	0.293		
Industry						
Education, Health & Social Work	0.043	0.204	0.114	0.318	0.082	0.275
Manufacturing	0.077	0.266	0.142	0.350	0.152	0.359
Mining	0.100	0.300	0.058	0.233	0.082	0.275
Commerce	0.146	0.353	0.430	0.497	0.463	0.499
Firm Size						
< 10	0.773	0.419	0.827	0.379	0.880	0.326
> 10	0.227	0.419	0.170	0.376	0.153	0.360
Other						
Location (Urban=1)	0.631	0.483	0.711	0.453	0.680	0.467
Age	32.80	10.62	31.83	10.71	33.18	9.204
Age squared	1189	799.5	1127	793.5	1185	653.9
Marital status (Married=1)	0.428	0.495	0.356	0.479	0.496	0.501
Tenure	6.862	6.798	4.074	5.745	3.896	4.834
Children	0.042	0.201	0.584	0.493	0.957	0.204
N	1613		731		415	

Figure 4.1 presents the kernel density plot for men and women in the two sectors. In the formal sector, the distributions of male and female wages are, to a large extent, symmetric and identical. In the left tails, however, the log wage distribution of women has a fatter tail compared to that of men implying that relatively more women earn low wages compared to men. In the right tails, on the other hand, there are more men earning higher wages compared to women. Additionally, the log wage distribution for women shows relatively less dispersion relative to that of men meaning there is less variation in the log wages of women compared to that of men. A similar pattern is observed in the informal sector except that the wage distribution of women is heavily skewed to the right. Overall, there are disproportionately more women earning low wages compared to men, but more men earning higher wages compared to women. Also, as we observe in the formal sample, the log wage distribution of women shows relatively less variation compared to the wage distribution of men.

The difference in log hourly wage without and with correction for sample selection, with 95% confidence intervals, between men and women at quantiles of their respective wage distributions is presented in Figures 4.2 and 4.3 below. In both cases, there is a clear upward trend of the gender log wage differential in the informal sector. In the formal sector, however, the unadjusted (conditional) log wage differential is negative in the lower quantiles but turns positive, and increases along the wage distribution. The implication is that, compared to men, women earn higher in the lower quantiles but this advantage vanishes and turns negative as we move along the wage distribution.

Finally, we also observe that, compared to the formal sector, the values of the raw wage differential are larger in the informal sector implying that there may be higher degrees of wage discrimination in the informal sector compared to the formal.

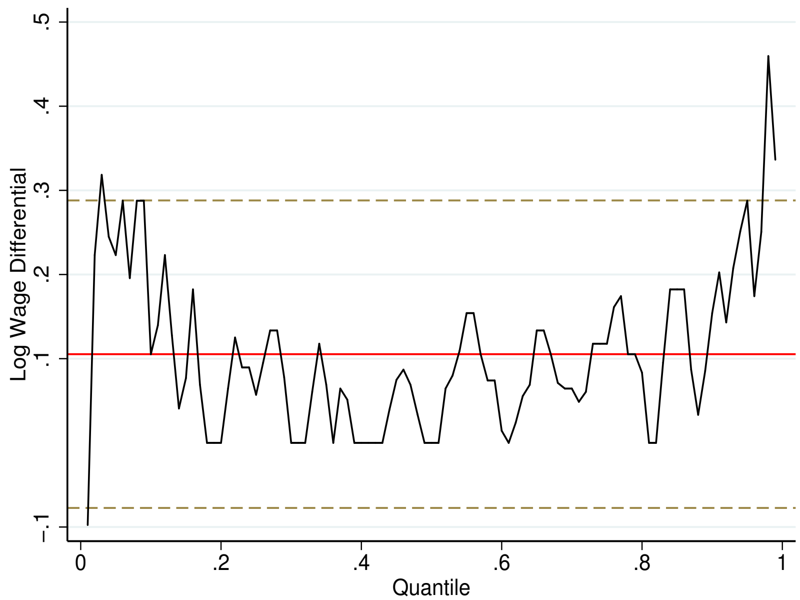
Figure 4.1: Kernel density of hourly earnings



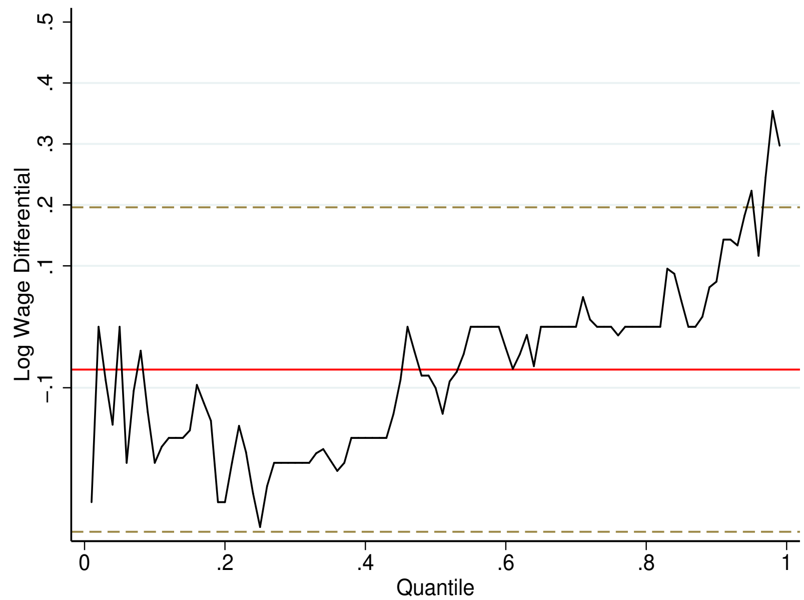
(a) Formal sector

(b) Informal sector

Figure 4.2: Formal sector log hourly wage gap

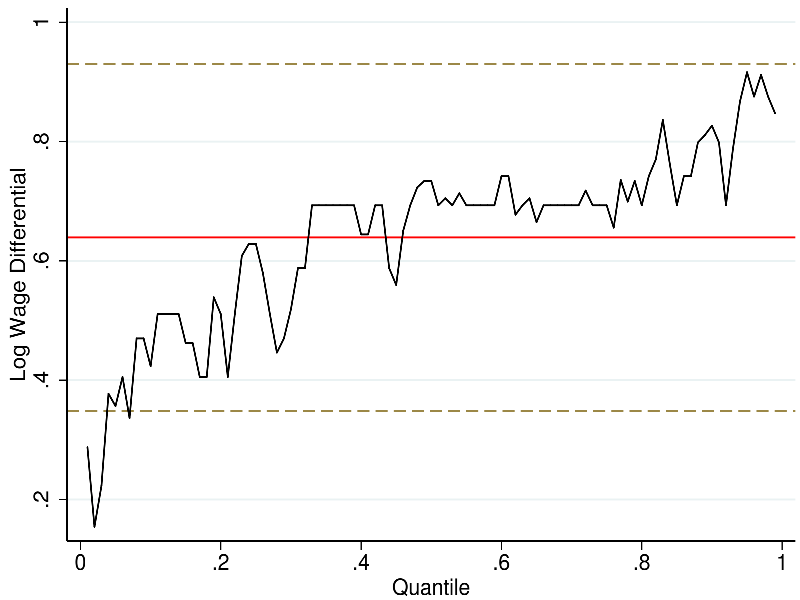


(a) Unadjusted

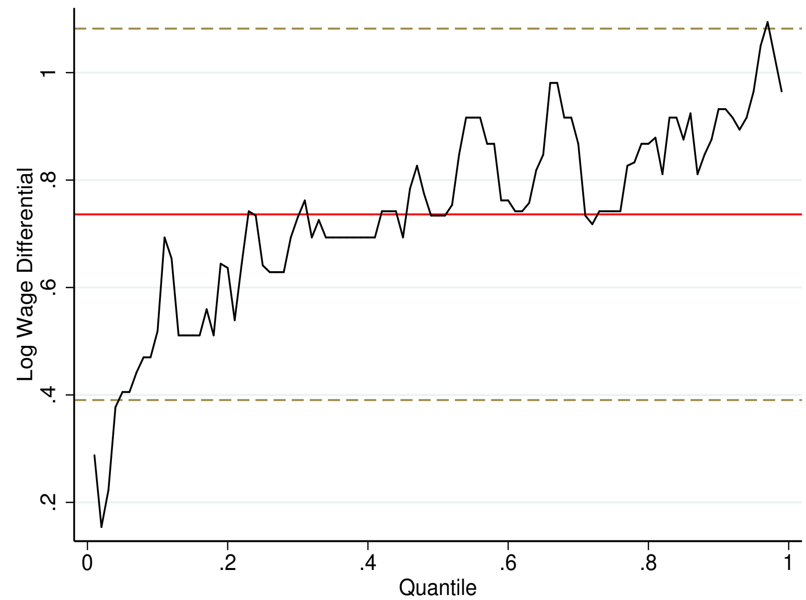


(b) Selection-adjusted

Figure 4.3: Informal sector log hourly wage gap



(a) Unadjusted



(b) Selection-adjusted

4.3.4 Determinants of the wage function

Previous empirical studies have adopted the human capital model as the theoretical basis for estimating the earning function [Becker, 1975; Mincer, 1974]. We make the same assumption here. The human capital model of wage determination posits that the structure of wages and earnings is mainly determined by a workers human capital - the stock of knowledge or characteristics that contributes to a worker's productivity. The level of education and work experience serve as proxies for measuring a worker's human capital. Education is measured by the highest level of education attained. In our sample, some 14% of men in the informal sector have completed at least primary education compared to 16% of women and 23% in the selected sample (see Table 4.2). 61% of men have completed secondary education compared to 49% of women and 45% in the selected sample. 7% of men compared to 9% of women have completed post-secondary education.

After accounting for selection the number of women with post-secondary education drops to zero meaning that relatively less educated women are more likely to work in the informal sector. On the whole, some 74% of men, compared to 68% of women, have completed either a primary or secondary education, and the typical informal sector worker has completed at least a secondary education. In the formal sector, 5% of men, compared to 2% of women, have completed some primary education (see Table 4.1). 39% of men, compared to 23% of women, have completed secondary education, and 55% of men against 74% of women have completed post-secondary education. This concentration of high skilled women in the formal sector may help explain why the raw gender wage gap is relatively smaller across the earnings distribution in the sector.

There are also some notable differences in education in the two samples. First, compared to the informal sector, there are fewer women with primary education compared to men, and second there are more women than men with post-secondary education. Overall, the typical formal sector worker is more educated than their informal counterpart –the average formal sector worker has completed a post-secondary education compared to primary education for the average informal sector worker. The two sectors seem to be segregated by education, with the formal sector attracting workers with more years of education, and low educated workers concentrated in the informal sector. The tenure variable is the proxy for worker experience, and it is measured in years. While in the formal sector women have more years of experience than men, the opposite is the case in the informal sector. Specifically, the average female worker in the formal sector has 11 years of experience compared to 7 years for the average man. In the informal sector, the numbers are 7 years for the average male worker and 4 years for the average woman.

Additional variables included in the wage function are dummies for industry of employment, firm size, region, marital status, children and age (in years). In the informal sector, women are concentrated in commerce with 46% employed there. This contrasts with only 15% for men. In the formal sector, on the other hand, larger proportions of both men (38%) and women (66%) are concentrated in the education, health and social work industry. The age distribution in the sample is similar across gender in the informal sector, but in the formal sector the average woman (38 years) is significantly older than the average man (33 years). We also see that more women (88%) than men (77%) in the informal sector work in firms with less than ten (10) employees. In the formal sector, on the other hand, fewer women (22%) compared to men (28%) work in firms with less than ten (10) employees. Overall, the average working woman in the informal sector, compared to the average man in same sector, is more likely to be working in commerce, has less work experience and more likely to be married with children. In the formal sector, however, the average working woman, compared to the average man, is more likely to be older, working in the education, health and social work industry and with more work experience.

4.4 Econometric Methodology

4.4.1 Quantile regression

Majority of previous discrimination literature has employed the canonical Oaxaca-Blinder (OB) (1973) or a variant of it in estimating the gender wage gap. Two key disadvantages of the OB estimator are that it can be inefficient if the assumption of normality is violated, and secondly the estimates can be biased in the presence of outliers. Quantile regression, on the other hand, is more robust to non-normal errors and able to capture heterogeneous effects, thus allowing for a richer characterization of the data. Due to this advantage, we employ the quantile regression technique in estimating the wage functions. We employ the quantile regression technique originally proposed by Koenker and Bassett [1978] to estimate the conditional earnings functions for men and women as:

$$y_i = x_i\beta_\theta + \mu_{\theta i}; \qquad Q_\theta(y_i|x_i) = x_i\beta_\theta \tag{4.1}$$

where y is the natural log of hourly wage, x is a vector of covariates, $i = 1 \ldots n$ indexes individuals, β is the vector of quantile regression coefficients and μ is the random error of an unknown distribution that should satisfy the restriction that $Q_\theta(\mu_{\theta i}|x_i) = 0$. The expression $Q_\theta(y_i|x_i)$ denotes the θ^{th} conditional quantile of y_i condition on the vector of covariates, x_i, with $0 < \theta < 1$. For a given quantile, θ, the quantile estimator,

β_θ, solves the minimization problem;

$$\beta_\theta = arg\,min_\beta \left[\sum_{i=y_i>x_i\beta} \theta|y_i - x_i\beta| + \sum_{i=y_i<x_i\beta} (1-\theta)|y_i - x_i\beta| \right] \tag{4.2}$$

Where $\theta|y_i - x_i|$ and $(1-\theta)|y_i - x_i\beta|$ represents the asymmetric penalties for under prediction and over prediction respectively.

4.4.2 *Wage gap decomposition*

Once the wage functions are estimated using quantile regression, we can then decompose the differences in the distribution to portions due to characteristics and coefficient. However, before we can proceed with the decomposition we need an estimate of the counterfactual wage function –that is, how much women would have earned if they were paid according to the wage function of men. This is analogous to estimating the 'treatment effect' of paying women according to the wage structure of men. This will imply constructing the unconditional wage distribution that females would have earned if they were males. We cannot proceed on the basis of the quantile wage functions because they are conditioned on the covariates.

A key complication with the quantile wage function is that because the law of iterated expectations does not apply to quantiles, the expected value of the conditional quantiles do not equate the unconditional quantiles. That is: $Q_\theta(y) \neq E[Q_\theta(y|x)]$, where $Q_\theta(y)$ is the unconditional wage distribution at the θ^{th} quantile, and $Q_\theta(y|x)$ is the corresponding conditional distribution. Melly [2005] proposes a technique to overcome this problem. This involves, first estimating the whole conditional distribution by quantile regression, and then estimating the unconditional distribution by integrating over the full set of the covariates as follows:

$$\theta = F_y(Q_\theta) = E[F_{y|x}(Q_\theta(y|x))] = \int F_{y|x}(Q_\theta(y|x))dF_z(x) \tag{4.3}$$

$F_y(Q_\theta)$ denotes the unconditional cumulative distribution of wages at the θ^{th} quantile. This (unconditional distribution) is then inverted to obtain the unconditional quantiles of interest.

Using this framework, the counterfactual wage distribution, $Q_\theta(y_m|x_f)$, can then be estimated and the wage gap decomposed as follows:

$$\Delta\theta = [Q_\theta(y_f|x_f) - Q_\theta(y_m|x_f)] + [Q_\theta(y_m|x_f) - Q_\theta(y_m|x_m)] \tag{4.4}$$

$Q_\theta(y_m|x_f)$ is the counterfactual wage distribution at the θ^{th} quantile. The first component on the right hand side represents difference due to coefficients and the second term measure the differences due to labour market characteristics (characteristics effect).

4.4.3 Selectivity Bias

The result of the decomposition may suffer from possible endogeneity bias. Men and women have different reasons that bring them into the labor market and different reasons for choosing to work in a particular sector. Presence of selection bias implies that the decomposition terms are not properly identified. Our results may suffer from possible selectivity bias which needs to be corrected. Previous empirical studies such as Albrecht et al. [2003]; Chzhen and Mumford [2011]; Nicodemo [2009] employed the semi-parametric estimator proposed by Buchinsky [1998] for selection correction in quantile regression. This procedure involves estimating a power series approximation of the selection term using the single-index method proposed by Ichimura [1993]. This term is then included in the quantile regression to account for selection. Consistency of the Buchinsky [1998] estimator depends on the assumption of conditional independence given the selection probability. That is, the error terms should be independent of the regressors in the presence of selectivity bias. In a recent publication, Huber and Melly [2011], however, show that in the presence of sample selection the conditional independence assumption is violated leading to inconsistent estimates.

We therefore apply a variant of the so called 'identification-at-infinity' [Chamberlain, 1986; Heckman, 1990] strategy to correct for the non-random participation of women in wage employment. The idea behind this procedure is to restrict the female sample (in both sectors) to a group for whom the choice to select into wage employment is not affected by the error term. We first run a probit model of participation for each sector using a sample of all women in the labour force then predict their probability of participation in employment, and then restrict the sample to only those whose predicted probability of participation is greater than 85% [Lindqvist and Vestman, 2011]. This selected sample is then used to run the decomposition regression again. However, because the decision to participate in wage employment is also conditioned on the decision to participate in the labour market in the first place we have a problem of double selection that needs to be jointly corrected. The two-step selection problem can be formally specified as follows;

Participation decision

$$Z_i' = w_i\alpha + \mu_{1i} \tag{4.5}$$

Where Z_i' is a latent variable that denotes the utility that a woman derives from participating in the labour force, w_i is a vector of covariates and μ_{1i} is the error term

$\sim N(0,1)$. With indicator variable;

$$Z_i = \begin{cases} 1 & \text{if } Z_i' > 0 \\ 0 & \text{if otherwise} \end{cases} \tag{4.6}$$

such that a woman participates in the labour force if $Z_i > 0$. The decision to participate in the labour force is followed by a choice of working in the formal (informal) sector. This decision is specified as follows:

Employment decision

$$S_i' = c_i\delta + \mu_{2i} \tag{4.7}$$

S_i' is a latent variable that represents the utility a woman derives working in the formal (informal) sector, c_i is a vector of covariates and μ_{2i} is the error term $\sim N(0,1)$. Employment in the formal (informal) sector is observed if;

$$S_i = \begin{cases} 1 & \text{if } S_i' > 0 \\ 0 & \text{if otherwise} \end{cases} \tag{4.8}$$

If the unmeasured factors influencing the error term μ_{1i} are uncorrelated with the factors influencing μ_{2i} then estimating a bivariate probit model will yield unbiased estimates. However, we have reason to suspect that this correlation might not be zero because there could be certain unobserved factors, that we haven't accounted for, that may jointly determine women's participation in the labour force and formal (informal) employment. To overcome this problem, Heckman [1979] proposes a selection correction procedure that allows for the possible correlation of the error terms. The technique was extended by Van de Ven and Van Praag [1981] to cover binary choice models. This technique known in the literature as '*heckprobit*' is employed in estimating the selection equations. Table 4.3 provides the results of the '*heckprobit*' estimation of the determinants of female participation in the labour force (column 2) and formal wage employment (column 1). The results (column 2) indicate that, unsurprisingly, education is the key driver of female participation in the labour force –women with higher education are more likely to participate in the labour force compared to women with no education. Living in an urban area is also positively correlated with the likelihood of participation in the labour force while women who live with their spouses are less likely to participate in the labour force. Uptake of formal employment is also driven by human capital factors as the results (in column 1) shows. Specifically, women with higher levels of education are more likely to take up formal employment. This result is expected given that highly educated women are more likely to have higher reservation wages and thus less likely to take up employment in the informal sector where traditionally wages are low compared to the formal sector.

Table 4.3: Incidence of formal wage employment among women

Variable	Formal Employment	Labour Force Participation
Constant	-4.261 (-3.48)***	-2.444 (-3.67)***
Education *(ref: No school)*		
Primary	0.232 (0.74)	0.329 (4.89)***
Secondary	1.027 (4.27)***	0.551 (8.58)***
Post-secondary	2.887 (6.56)***	1.884 (21.72)***
Other		
Location (Urban=1)	0.032 (0.08)	0.536 (10.84)***
Age	0.066 (0.99)	-0.021 (-1.08)
Age squared	-0.007 (-0.81)	0.0002 (0.74)
Marital status (Married=1)	-0.057 (-0.35)	
Children	0.402 (0.50)	-0.228 (-1.01)
Spouse live in HH		-0.131 (2.25)**
ρ	0.475 (0.25)	
N	9137	9137

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

The '*heckprobit*' results for participation in the informal sector are presented in Table 4.4. Again, we see that higher education is associated with a higher likelihood of participation in the labour force (see column 2). Women living in urban areas are more likely to participate in the labour force. Women with children and women who are heads of their households both have higher likelihood of labour force participation but married women are less likely to participate in the labour force. Just as we observed in the formal sector, uptake of informal employment is driven mainly by level of education. Specifically, highly educated women are less likely to take up informal employment relative to women with less education. As argued earlier, this may be due to the fact that highly educated women may prefer formal employment because that sector traditionally pays higher wage compared to the informal sector. In both Tables 4.3 and 4.4 the Wald-test of independent equations fails to reject the null hypothesis at the 5% significance level indicating that the error terms of equations 4.5 and 4.7 are uncorrelated, and thus the participation wage employment decisions are independent.

Table 4.4: Incidence of informal wage employment among women

Variable	Informal Employment	Labour Force Participation
Constant	2.862 (4.29)*	2.521 (10.96)***
Education *(ref: No school)*		
Primary	-0.215 (-1.40)	0.209 (3.61)***
Secondary	-0.751 (-6.38)***	0.418 (9.07)***
Post-secondary	-2.241 (-8.17)***	1.689 (24.26)***
Other		
Location (Urban=1)	-0.256 (-0.73)	0.526 (12.92)***
Age	-0.034 (-1.10)	0.011 (0.60)
Age squared	0.0004 (1.09)	-0.0002 (-0.97)
Marital status (Married=1)		-0.163 (-2.98)**
Children	-0.184 (-0.41)	0.525 (4.39)***

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

. . . continued

Table 4.4: Incidence informal wage employment among women, *continued ...*

Variable	informal Employment	Labour Force Participation
Spouse live in HH	0.116 (0.82)	
Household Head		0.155 (2.87)**
ρ	-0.626 (0.36)	
N	10959	10959

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

4.5 Results

4.5.1 Quantile regression

Table 4.5 presents the results (without selection correction) of the conditional quantile regressions for log hourly wage for both men and women in the formal sector. The cumulative effect of age on log hourly wage is positive and significant for men at only the 25^{th} and 50^{th} quantiles, but this effect falls as we move along the wage distribution. For women, however, the effect of age on wages is insignificant. The returns to primary and secondary education for men are positive and u-shaped as we move along the wage distribution. However, the size of the returns is larger for men and women with post-secondary education relative to those with secondary education.

Similar pattern is observed for the return to post-secondary education for women. For secondary education, on the other hand, the return is significantly larger for higher earning women compared to women in the low earning women. Effect of tenure levels on wages of men is positive and significant but the magnitude is small. For women, on the other hand, this effect is only significant and positive for low earning women. Women working in manufacturing and electricity and utilities sectors earn significantly lower wages compared to other sectors, but this effect is only significant at the 50^{th} and 75^{th} quantiles. For men, compared to other sectors, working in mining and construction and finance and public administration is associated with higher wages and the size of this effect is larger in the higher quantiles. The effect of firm size on wages of both men and women is non-trivial. Specifically, men and women who work in firms with less than ten (10) employees earn significantly lower wages compared to those who work in firms with more than ten (10) employees. This is unsurprising because a firm's size is directly related to its productive capacity and so large firms, ceteris paribus, are able to pay higher wages compared to smaller firms.

After correcting for the non-random selection of women into formal employment, the results and their signs are not significantly different (see Table 4.6). However, the only variable consistently significant across all quantiles is post-secondary education. Like the last results, the returns to post-secondary education increases along the wage distribution. Women with children earn lower wages compared to those with no children. However, this effect is only significant at the 25^{th} and 50^{th} quantiles.

Table 4.5: Quantile earnings regression (Formal sector)

	Male			Female		
Variable	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
Constant	-0.689 (-1.42)	0.876 (2.12)**	1.303 (2.86)***	0.564 (0.82)	1.072 (2.11)**	1.356 (2.17)**
Education *(ref: No school)*						
Primary	0.223 (0.93)	0.025 (0.12)	0.097 (0.43)	-0.301 (-0.81)	-0.441 (-1.60)	0.326 (0.96)
Secondary	0.655 (3.04)***	0.460 (2.50)**	0.554 (2.73)***	0.296 (1.09)	0.373 (1.85)*	0.639 (2.57)**
Post-secondary	1.318 (6.11)***	1.112 (6.04)***	1.185 (5.83)***	1.060 (3.88)***	1.032 (5.09)***	1.271 (5.09)***
Industry						
Education, Health & Social Work	0.032 (0.49)	-0.024 (-0.43)	-0.119 (-1.92)*	0.063 (0.55)	-0.162 (-1.91)*	-0.219 (-2.11)**
Manufacturing	-0.076 (-0.74)	0.009 (0.10)	0.058 (0.59)	-0.270 (-1.20)	-0.803 (-4.79)***	-0.868 (-4.21)***
Mining & Construction	0.327 (3.21)***	0.378 (4.34)***	0.399 (4.15)***	0.700 (1.71)*	0.115 (0.38)	-0.006 (-0.02)
Public Admin & Finance	0.235 (3.07)***	0.271 (4.14)***	0.307 (4.24)***	0.094 (0.64)	-0.015 (-0.14)	0.036 (0.27)
Electricity & Utilities	-0.442 (-1.59)	-0.105 (-0.45)	0.093 (0.36)	-0.183 (-0.40)	-0.851 (-2.51)**	-0.808 (-1.94)*
Firm Size						
< 10	-0.150 (-2.70)***	-0.110 (-2.32)**	-0.188 (-3.57)***	-0.161 (-1.67)*	-0.220 (-3.08)***	-0.158 (-1.79)*
Other						
Location (Urban=1)	0.064 (1.09)	0.095 (1.90)*	0.093 (1.69)*	0.071 (0.71)	0.029 (0.39)	0.086 (0.94)
Age	0.090 (4.12)***	0.040 (2.17)**	0.032 (1.57)	0.032 (0.91)	0.038 (1.48)	0.019 (0.60)
Age squared	-0.001 (-4.14)***	-0.001 (-2.28)**	-0.0004 (-1.64)	-0.0005 (-1.06)	-0.0005 (-1.42)	-0.0001 (-0.25)
Marital status (Married=1)	0.067 (1.03)	0.065 (1.16)	0.111 (1.80)*	0.104 (1.22)	0.120 (1.89)*	0.006 (0.08)
Tenure	0.009 (2.65)***	0.010 (3.21)***	0.010 (2.93)***	0.019 (3.11)***	0.012 (2.66)***	0.005 (0.97)
Children	0.061 (0.47)	0.089 (0.81)	-0.034 (-0.28)	0.042 (0.42)	-0.089 (-1.21)	0.047 (0.52)

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

. . . continued

Table 4.5: Quantile earnings regression (Formal sector), *continued ...*

	Male			Female		
Variable	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
N	1114	1114	1114	545	545	545

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Table 4.6: Selection-corrected quantile earnings regression (Formal sector)

	Male			Female		
Variable	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
Constant	-0.689 (-1.42)	0.876 (2.12)**	1.303 (2.86)***	0.935 (1.46)	1.611 (2.58)**	1.314 (1.61)
Education *(ref: No school)*						
Primary	0.223 (0.93)	0.025 (0.12)	0.097 (0.43)	-0.580 (-1.50)	-0.468 (-1.24)	0.352 (0.71)
Secondary	0.655 (3.04)***	0.460 (2.50)**	0.554 (2.73)***	0.231 (0.75)	0.407 (1.35)	0.722 (1.83)*
Post-secondary	1.318 (6.11)***	1.112 (6.04)***	1.185 (5.83)***	0.774 (2.51)**	0.940 (3.12)***	1.215 (3.08)***
Industry						
Education, Health & Social Work	0.032 (0.49)	-0.024 (-0.43)	-0.119 (-1.92)*	0.097 (1.04)	-0.080 (-0.87)	-0.123 (-1.03)
Manufacturing	-0.076 (-0.74)	0.009 (0.10)	0.058 (0.59)	-0.498 (-2.19)**	-0.692 (-3.12)***	0.047 (0.16)
Mining & Construction	0.327 (3.21)***	0.378 (4.34)***	0.399 (4.15)***	0.569 (1.70)*	0.383 (1.18)	0.125 (0.29)
Public Admin & Finance	0.235 (3.07)***	0.271 (4.14)***	0.307 (4.24)***	0.151 (1.28)	0.066 (0.57)	0.208 (1.38)
Electricity & Utilities	-0.442 (-1.59)	-0.105 (-0.45)	0.093 (0.36)	-0.375 (-0.98)	-0.408 (-1.10)	-0.644 (-1.32)
Firm Size						
< 10	-0.150 (-2.70)***	-0.110 (-2.32)**	-0.188 (-3.57)***	-0.203 (-2.57)**	-0.159 (-2.07)**	-0.124 (-1.23)
Other						
Location (Urban=1)	0.064 (1.09)	0.095 (1.90)*	0.093 (1.69)*	0.086 (1.02)	0.018 (0.22)	0.120 (1.11)
Age	0.090 (4.12)***	0.040 (2.17)**	0.0322 (1.57)	0.048 (1.63)	0.021 (0.72)	0.024 (0.64)
Age squared	-0.001 (-4.14)***	-0.001 (-2.28)**	-0.0004 (-1.64)	-0.0007 (-1.78)*	-0.0003 (-0.77)	-0.0002 (-0.34)
Marital status (Married=1)	0.067 (1.03)	0.065 (1.16)	0.111 (1.80)*	0.096 (1.38)	0.113 (1.67)*	-0.003 (-0.04)
Tenure	0.009 (2.65)***	0.010 (3.21)***	0.010 (2.93)***	0.015 (2.97)***	0.011 (2.14)**	0.006 (0.90)
Children	0.061 (0.47)	0.089 (0.81)	-0.034 (-0.28)	-0.385 (-4.47)***	-0.224 (-2.66)***	-0.105 (-0.95)

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

...continued

Table 4.6: Selection-corrected quantile earnings regression (Formal sector), *continued ...*

	Male			Female		
Variable	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
N	1114	1114	1114	419	419	419

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

The conditional quantile earning regression results (without selection-correction) for the informal sector is presented in Table 4.7. For men, the cumulative effect of age on wages is positive at all the quantiles. As expected, younger men earn significantly higher wages than older men and this effect is stronger for low earning men. For women, however, this effect is only significant at the 50^{th} quantile. At all the quantiles, married men earn significantly higher wages relative to single men. Wages of married and single women, on the other hand, are not statistically different.

For the human capital variables, the only consistently significant factor in wage determination in this sector is post-secondary education. Both men and women with post-secondary education earn significantly higher wages relative to those with no education. However, the returns for women are higher compared to that of men. This result may be because women with higher education are relatively scarce in the informal sector, and so employers may have to offer a significantly higher wages to attract them to the sector. Both men and women working in the education, health and social work sector their wages are significantly lower compared to those working in other sectors. This story, however, contrasts with returns in the mining sector. Men working in mining earn significantly higher wages compared to other industries and the returns are highest for low earning men, but decreases along the wage distribution. Women working in mining also earn higher wages relative to women in other occupations. In contrast with men, the returns of women working in mining tend to follow an inverted u-shape pattern along the distribution; it increases through the 50^{th} quantile and falls by the 75^{th} quantile. Unsurprisingly, living in an urban area is positively correlated with wages at all the quantiles. And as observed in the formal sector, workers working in smaller size firms earn significantly lower wages compared to those that work in relatively larger firms.

Table 4.7: Quantile earnings regression (Informal sector)

	Male			Female		
Variable	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
Constant	-0.445 (-1.38)	0.363 (1.20)	1.247 (3.22)***	0.570 (1.22)	0.278 (0.70)	0.863 (2.02)**
Education *(ref: No school)*						
Primary	0.009 (0.09)	-0.176 (-1.81)*	-0.190 (-1.53)	0.015 (0.11)	-0.045 (-0.42)	0.266 (2.25)**
Secondary	0.186 (2.29)**	0.036 (0.48)	-0.062 (-0.63)	0.146 (1.40)	0.150 (1.70)*	0.362 (3.80)***
Post-secondary	0.489 (3.70)***	0.220 (1.77)*	0.393 (2.47)**	0.779 (4.46)***	0.740 (5.00)***	0.914 (5.73)***
Industry						
Education, Health & Social Work	-0.315 (-1.98)**	-0.283 (-1.90)*	-0.625 (-3.28)***	-0.300 (-1.95)*	-0.343 (-2.64)***	-0.406 (-2.89)***
Manufacturing	-0.051 (-0.48)	-0.079 (-0.78)	-0.108 (-0.85)	-0.370 (-2.82)***	-0.135 (-1.21)	-0.234 (-1.95)*
Mining	0.823 (7.66)***	0.816 (8.07)***	0.645 (5.00)***	0.724 (3.32)***	0.760 (4.11)***	0.720 (3.62)***
Commerce	-0.065 (-0.77)	-0.041 (-0.52)	-0.043 (-0.42)	-0.152 (-1.50)	-0.030 (-0.34)	-0.075 (-0.81)
Firm Size						
< 10	-0.084 (-1.17)	-0.183 (-2.73)***	-0.175 (-2.04)**	-0.234 (-2.05)**	-0.137 (-1.42)	-0.251 (-2.41)**
Other						
Location (Urban=1)	0.341 (5.41)***	0.298 (5.04)***	0.206 (2.73)***	0.200 (2.03)**	0.264 (3.17)***	0.311 (3.46)***
Age	0.077 (4.24)***	0.069 (4.02)***	0.057 (2.60)***	0.015 (0.55)	0.043 (1.86)*	0.023 (0.92)
Age squared	-0.001 (-4.45)***	-0.001 (-3.81)***	-0.001 (-2.59)***	-0.0003 (-0.73)	-0.001 (-1.91)*	-0.0003 (-0.78)
Marital status (Married=1)	0.154 (2.25)**	0.152 (2.36)**	0.221 (2.69)***	0.061 (0.65)	-0.023 (-0.29)	-0.006 (-0.07)
Tenure	0.013 (2.49)**	0.005 (1.04)	0.004 (0.68)	0.015 (1.79)*	0.025 (3.55)***	0.020 (2.64)***
Children	-0.008 (-0.06)	0.020 (0.15)	-0.026 (-0.16)	-0.114 (-1.10)	-0.169 (-1.92)*	-0.016 (-0.17)
N	1206	1206	1206	545	545	545

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

After correcting for sample selection for women most of the estimates lose their significance though the signs of the estimates are not significantly different (see Table 4.8). The predicted probability of participation in the informal sector for all women with post-secondary education was less than 0.85, and so they were excluded from the selected sample. This suggests that women in the selected sample are less educated than women in the population at large. The effect of tenure levels at all the quantiles is no longer significant for women. The return to working in the mining sector is increasing along the distribution. Given that all the human capital variables are not significant in determining wages in this sector, this may suggest that wage determination in the informal sector may be driven by some social, institutional and cultural factors such as 'relationship to the employer', ethnicity, religion, whether one is a household head etc. However, due to data limitations we were unable to test this hypothesis.

Table 4.8: Selection-corrected quantile earnings regression (Informal sector)

	Male			Female		
Variable	25^{th}	50^{th}	75^{th}	25^{th}	50^{th}	75^{th}
Constant	-0.445 (-1.38)	0.363 (1.20)	1.247 (3.22)***	0.422 (0.76)	0.679 (1.57)	0.710 (1.14)
Education *(ref: No school)*						
Primary	0.009 (0.09)	-0.176 (-1.81)*	-0.190 (-1.53)	-0.120 (-0.93)	-0.174 (-1.74)*	0.160 (1.12)
Secondary	0.186 (2.29)**	0.036 (0.48)	-0.062 (-0.63)	0.052 (0.46)	0.004 (0.04)	0.195 (1.55)
Post-secondary	0.489 (3.70)***	0.220 (1.77)*	0.393 (2.47)**			
Industry						
Education, Health & Social Work	-0.315 (-1.98)**	-0.283 (-1.90)*	-0.625 (-3.28)***	-0.316 (-1.61)	-0.102 (-0.67)	-0.184 (-0.84)
Manufacturing	-0.051 (-0.48)	-0.079 (-0.78)	-0.108 (-0.85)	-0.333 (-2.12)**	-0.087 (-0.71)	0.072 (0.41)
Mining	0.823 (7.66)***	0.816 (8.07)***	0.645 (5.00)***	0.792 (3.99)***	0.820 (5.31)***	0.828 (3.74)***
Commerce	-0.065 (-0.77)	-0.041 (-0.52)	-0.043 (-0.42)	-0.138 (-1.12)	0.082 (0.86)	0.142 (1.04)
Firm Size						
< 10	-0.084 (-1.17)	-0.183 (-2.73)***	-0.175 (-2.04)**	-0.215 (-1.44)	-0.148 (-1.28)	-0.130 (-0.78)
Other						
Location (Urban=1)	0.341 (5.41)***	0.298 (5.04)***	0.206 (2.73)***	0.192 (1.76)*	0.141 (1.67)*	0.173 (1.43)
Age	0.077 (4.24)***	0.069 (4.02)***	0.057 (2.60)***	0.024 (0.73)	0.028 (1.07)	0.020 (0.52)
Age squared	-0.001 (-4.45)***	-0.001 (-3.81)***	-0.001 (-2.59)***	-0.0004 (-0.77)	-0.0004 (-1.09)	-0.0002 (-0.46)
Marital status (Married=1)	0.154 (2.25)**	0.152 (2.36)**	0.221 (2.69)***	0.115 (1.19)	0.030 (0.40)	-0.009 (-0.08)
Tenure	0.013 (2.49)**	0.005 (1.04)	0.004 (0.68)	0.005 (0.51)	0.006 (0.79)	0.012 (0.97)
Children	-0.008 (-0.06)	0.020 (0.15)	-0.026 (-0.16)	-0.103 (-0.39)	-0.136 (-0.67)	0.174 (0.60)
N	1206	1206	1206	415	415	415

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

4.5.2 *Quantile Decomposition*

Finally, we present the results of the decomposition of differences in distribution. The formal sector decomposition results are presented in Table 4.9. The result of the decomposition (not adjusted for selection) indicates that the conditional wage gap is negative and increases significantly between the 50^{th} and 75^{th} quantiles, but is insignificant at the 25^{th} quantile. Specifically, at the 50^{th} quantile, the average woman earns 6.8% $[(e^{(-0.071)} - 1) * 100]$ less than an average man with similar characteristics. At the 75^{th} quantile, this gap jumps to 10%. Difference due to characteristics, or the earning differential due to skill or endowment, is positive at all the quantiles but only significant at the 25^{th} and 50^{th} quantile, indicating that, relative to an average man, the average woman is more 'skillful'. Differences due to coefficients, or the unexplained part of the wage gap, are negative at all quantiles, but largest at the lower and upper quantiles. Taken together, the conditional wage gap is relatively smaller compared to the differential due to endowments because the positive difference due to endowments has a dampening effect on the conditional wage gap.

Table 4.9: Decomposition of differences in distribution (Formal sector)

Quantile	Without selection-correction	With selection-correction
25^{th}		
Raw difference	-0.049 (-0.89)	0.217 (4.61)***
Characteristics	0.085 (0.94)**	0.197 (2.31)***
Coefficients	-0.134 (-3.95)***	0.020 (0.63)
50^{th}		
Raw difference	-0.071 (-1.96)**	0.095 (3.13)***
Characteristics	0.055 (0.92)**	0.153 (2.15)***
Coefficients	-0.125 (-4.65)***	-0.058 (-2.21)**
75^{th}		
Raw difference	-0.106 (-3.72)***	-0.004 (-0.12)

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

. . . continued

Table 4.9: Decomposition of differences in distribution (Formal sector), *continued ...*

Quantile	Without selection-correction	With selection-correction
Characteristics	0.007 (0.11)	0.073 (1.14)***
Coefficients	-0.113 (-19.57)***	-0.077 (-3.28)***
Number of Observations	Men = 1114 Women = 515	Men = 1114 Women = 419

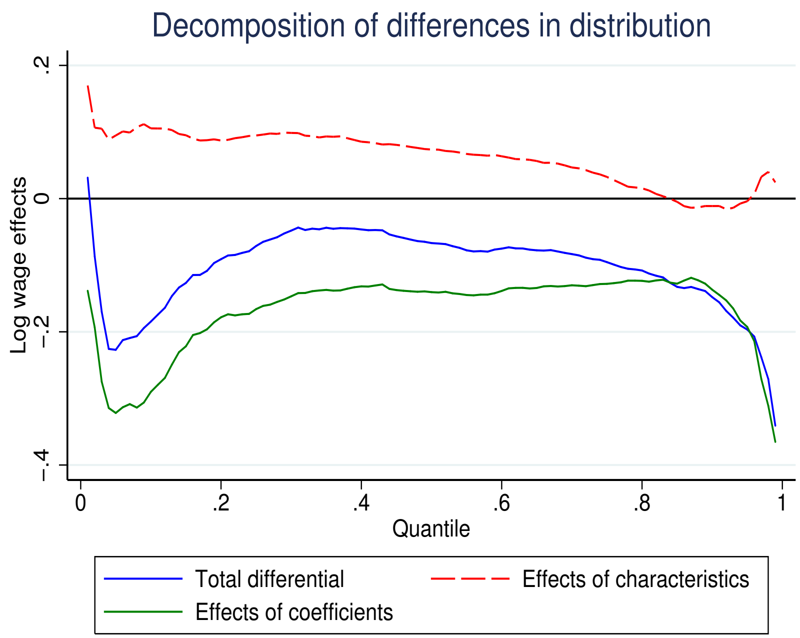
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

The graph of how the (unadjusted) wage gap varies along the entire wage distribution is produced in Figure 4.4. Interestingly, at almost all the quantiles, the average woman is more skillful compared to the average man. The difference due to coefficient, or the unexplained portion of the wage gap, however, varies along the wage distribution. The absolute value of the conditional wage gap is greatest at the lower and upper quantiles, but relatively flat everywhere else.

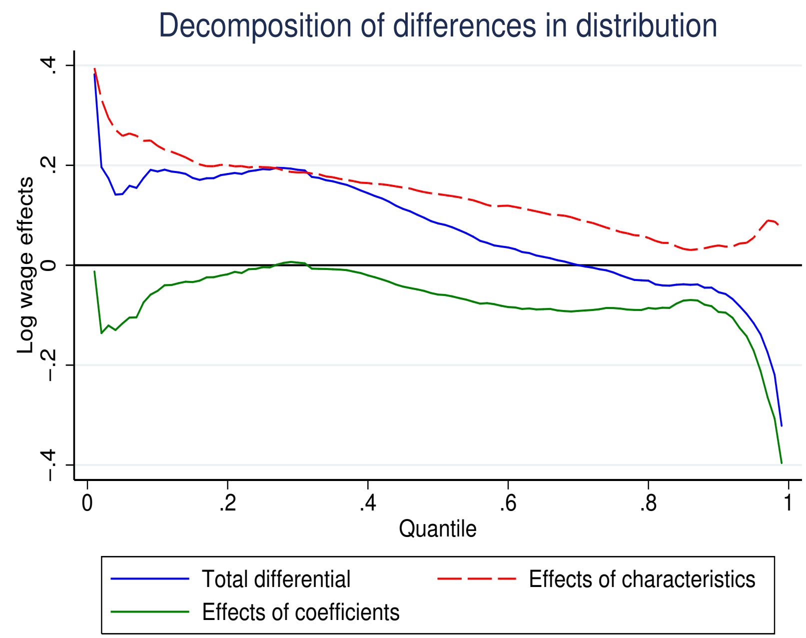
After correcting for sample selection, the nature of the wage gap changes significantly (see Table 4.9). The gap is positive and significant at the 25^{th} and 50^{th} quantiles, but insignificant at the 75^{th} quantile. Specifically it is 24% and 10% at the 25^{th} and 50^{th} quantiles, respectively. This implies that at the 25^{th} and 50^{th} quantiles, the average woman earns a higher wage compared to an average man with identical characteristics. To put this in perspective, at the 25^{th} quantile a man with average female characteristics, compared to the average woman, earns 76% per currency unit and 90% per currency unit at the 50^{th} quantile. This positive wage gap is, however, insignificant at the 75^{th} quantile. The trend over the entire wage distribution is presented in Figure 4.4. The conditional wage gap is positive and falling up to the 70^{th} quantile, but turns negative afterwards. The differential due to endowment is positive for all women in the sector, even though it is falling along the wage distribution. Even though, in general, there is a positive wage gap we note that the differential due to coefficient is negative below the 20^{th} and above the 40^{th} quantiles (see Figure 4.5).

This means that the conditional wage gap is positive mainly because difference due to endowments explains more than the total observed wage gap.

Figure 4.4: Formal sector decomposition plot

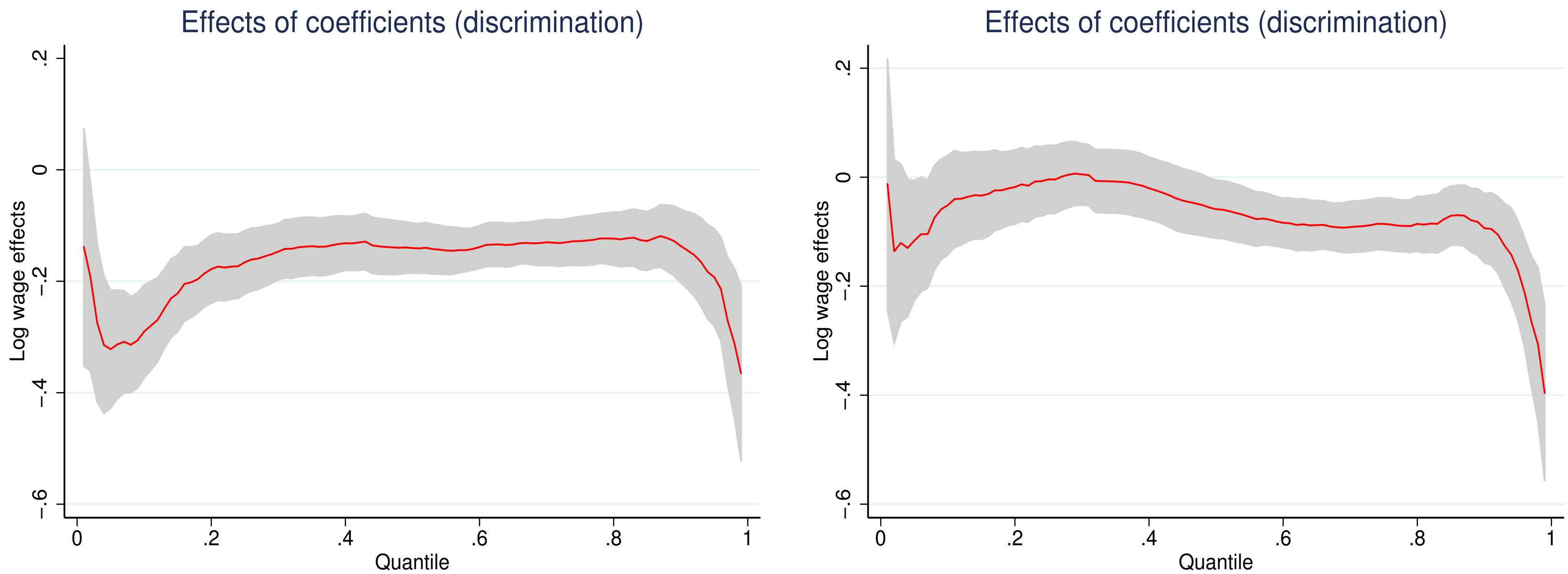


(a) Unadjusted



(b) Selection-adjusted

Figure 4.5: Formal sector discrimination plot



(a) Unadjusted

(b) Selection-adjusted

In contrast, the decomposition results of the informal sector tell a different story. We observe significant negative conditional wage gaps at all quantiles; both before and after selection correction (see Table 4.10). The results (not adjusted for selection) indicate that the conditional wage gap is 44% at the 25^{th} quantile, 51% at the 50^{th} and 52% at the 75^{th} quantile, indicating that the gap is increasing along the quantiles. At the 25^{th} quantile, differences due to endowment accounts for 18% of the observed wage gap while the remaining 82% is due to differences in coefficient. At the 50^{th} quantile, 17% of the conditional wage gap is due to differences in endowment and 83% due to coefficients (or unexplained) while at the 75^{th} quantile 18% is explained by differences in endowment and 82% explained by differences coefficients.

A graph of how the wage gap varies along the entire wage distribution is presented in Figure 4.6. The absolute value of the conditional wage differential is increasing along the wage distribution. Along the distribution, the average man is better endowed relative to the average woman, and this difference in endowment explains about 0% to 19% of the observed conditional wage gap. Difference due to coefficient, or the unexplained portion of the wage gap, is the largest component of the observed wage gap and this (gap) increases significantly as we move along the wage distribution.

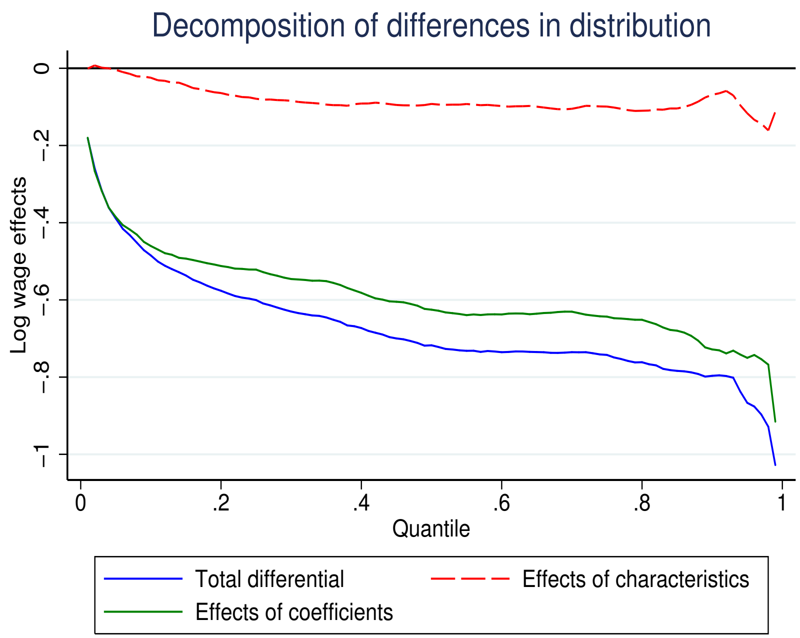
After accounting for selection of women into informal employment, the gap widens at all quantiles (see Table 4.10). The general pattern of the conditional wage gap still persist –increasing along the entire wage distribution (Figure 4.6). Differences in endowments do not explain much of the gap –6% at the 25^{th} quantile and 10% at the 75^{th} quantile. This means that the increase in the observed wage gap is driven mainly by increase in the proportion of the gap due to differences in coefficients (see Figure 4.7). This is unsurprising given that the quantile wage regression of the selected sample produced very few significant coefficients. We should, therefore, be cautious in explaining away the differences due to coefficients as discrimination. This is because, this portion of the conditional wage gap may also be capturing the effect of omitted and unobserved variables, and thus would be overestimating the size of discrimination.

Table 4.10: Decomposition of differences in distribution (Informal sector)

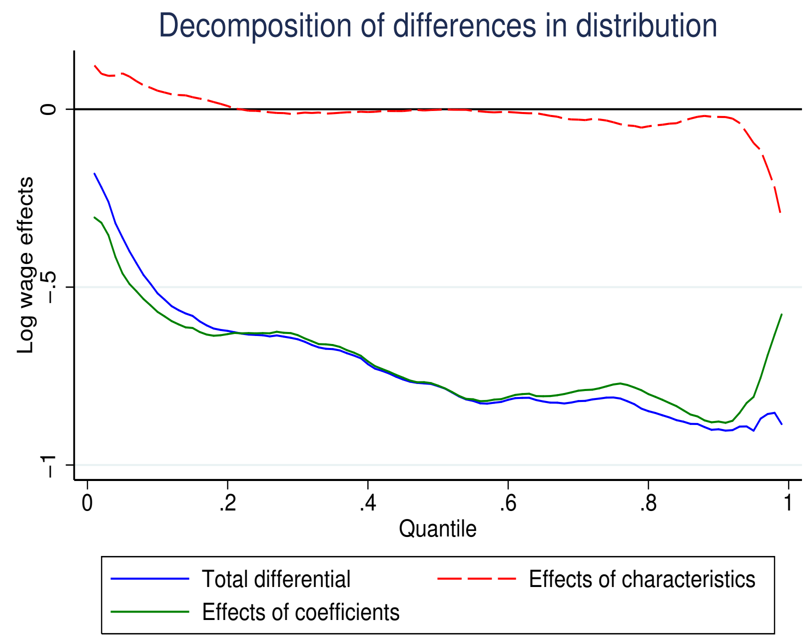
Quantile	Without selection-correction	With selection-correction
25^{th}		
Raw difference	-0.593 (-17.87)***	-0.635 (-17.68)***
Characteristics	-0.108 (-1.32)***	-0.041 (-0.29)*
Coefficients	-0.485 (-15.47)***	-0.594 (-24.57)***
50^{th}		
Raw difference	-0.711 (-23.14)***	-0.796 (-20.82)***
Characteristics	-0.119 (-1.50)***	-0.030 (-0.22)
Coefficients	-0.591 (-19.36)***	-0.765 (-33.00)***
75^{th}		
Raw difference	-0.741 (-18.24)***	-0.801 (-17.61)***
Characteristics	-0.128 (-1.53)***	-0.079 (-0.60)***
Coefficients	-0.613 (-19.57)***	-0.722 (-24.79)***
Number of	Men = 1205	Men = 1205
Observations	Women = 545	Women = 415

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Figure 4.6: Informal sector decomposition plot

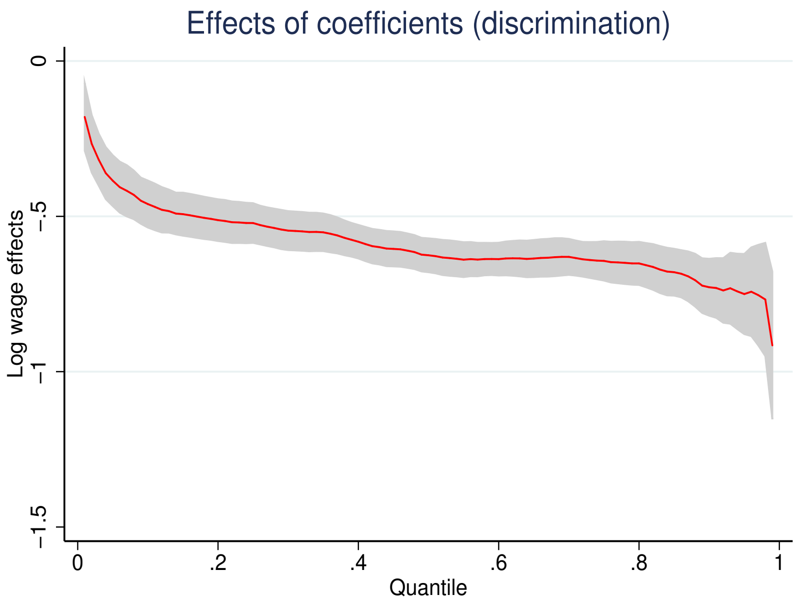


(a) Unadjusted

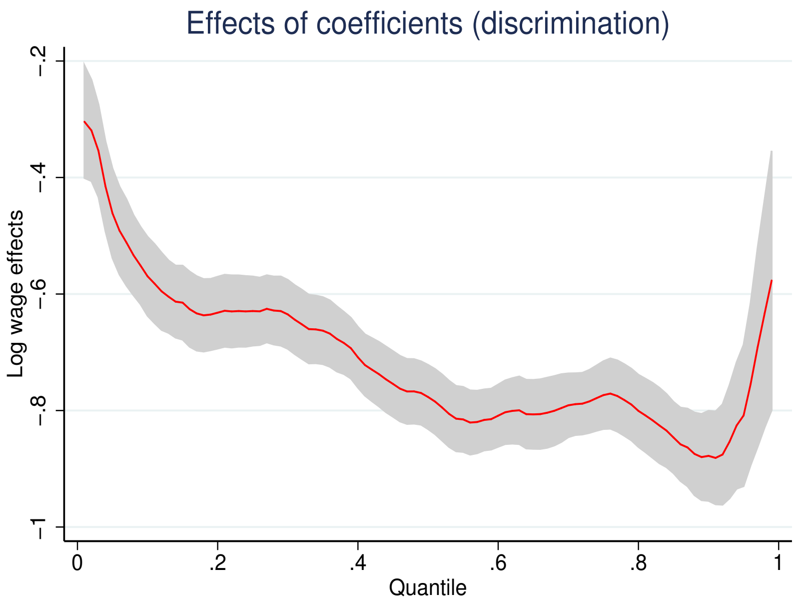


(b) Selection-adjusted

Figure 4.7: Informal sector discrimination plot



(a) Unadjusted



(b) Selection-adjusted

4.6 Conclusion

Using data from the sixth round of the Ghana Living Standard Survey 2012/2013 we examine the gender wage gaps in the formal and informal sectors. The empirical results show that, after accounting for observable characteristics, women in informal employment earn less relative to men; consistent with the result of Addai [2011]. This wage gap is smaller at lower quantiles, but increases along the wage distribution. In the formal sector, on the other hand, the wage gap is largest at the lower and upper quantiles, but relatively flat at the mid quantiles.

After correcting for positive selection of women into wage employment using an identification at infinity strategy, we found that though the trend of the gap is mostly unchanged in the informal sector, the magnitude of the gap increases at all quantiles. We note that differences in characteristics or the unexplained portion of the gap ranges between 94% at the 25^{th} quantile and 90% at the 75^{th} quantile. This may suggest the presence of discrimination in the informal sector, but we have to exercise caution in ascribing the entire unexplained portion to discrimination. This is because the conditional quantile wage regressions for the selected sample indicated that most of the observable individual characteristics such as level of education, tenure and age are not significant determinants of wages in the informal sector. This may suggest that the human capital theory of wage formation may not be a good theoretical basis for wage determination in the informal sector, but rather factors as such social norms or cultural factors may be more relevant. On this basis, the estimated size of 'discrimination' will, therefore, be conditioned on the weight of any excluded exogenous variables, as well as, other unobserved variables that may influence wages.

However, the sheer size of the unexplained portion of the wage gap alone suggests there is still room for discrimination after we account for the biases induced by omitted and unobserved variables. In the formal sector, on the other hand, the wage gap is positive and large at lower quantiles, but disappears at the top. This implies that, along the wage distribution, women earn higher wages than men but this gap disappears at the top. In contrast to the informal sector, almost the entire (conditional) wage differential is explained by differences in observable individual characteristics. In comparing the (conditional) wage gaps in the two sectors, we see that while there is a relatively larger gender wage gap in favor of men in the informal sector, in the formal sector, the wage gap favors women. This result may hint at the presence of segmentation in the Ghanaian labour market which may be explained by the roles of trade unions, collective bargaining and labour standards in the formal sector. In fact, empirical work by Blau and Kahn [2003] shows that gender wage differentials are significantly influenced by the overall distribution of wages in a country. In particular, the broader the area covered by collective negotiations –which generally leads to a reduction in the spread of wages– the less the gender wage gap. This may explain why the wage distribution for men and women in the formal sector is very identical. The implication of this result is that formalization may play a role in reducing the observed gender wage gap in the informal sector. Additionally, the observation of possible discrimination in both sectors implies that the Ghanaian labour market is not competitive.

We note key limitations to the generalization of this study. First is the possible overestimation of the size of discrimination in the informal sector due to the conflation of the effects of omitted and unobservable variables with the 'true' value of discrimination. Data limitation did not allow us to test the significance of other cultural and institutional variables in wage determination in the informal sector. Another issue is that selection correction using identification-at-infinity limits the results to the specific selected sample, and thus cannot be generalized to the entire population. However, given that both the results of the conditional quantile regressions with and without selection correction points towards a larger wage gap in the informal sector, relative to the formal, we can say with confidence that the wage gap in the informal sector is larger compared to the formal.

Finally, the use of cross-sectional data does not tell us how the two sectors interact and so we are unable to capture the dynamics of the wage gaps. Therefore, future research that uses panel data and a larger sample size may throw more light on the dynamics of the gender wage gap in both sectors.

5. SUMMARY AND CONCLUSIONS

The purpose of this thesis was to empirically analyze the nature of informal wage employment and how certain aspects of informality interact with health care and wage inequality. In this regard, the first essay explores heterogeneity in informal wage employment and tests the hypothesis of labour market segmentation in the Ghanaian labour market. The second essay examines the determinants of voluntary health insurance uptake by informal members of Ghana's National Health Insurance Scheme (NHIS) and how voluntary uptake of health insurance influences health care use. The final essay undertakes a comparative analysis of the gender wage gaps in the formal and informal sectors of the Ghanaian labour market.

The first essay attempts to address two main questions. First, is all wage employment in the informal sector voluntary? In other words, is the Ghanaian labour market segmented or competitive with regards to wage employment? And second, whether informal wage employment is heterogeneous, and, if it is, how many distinct divisions are there? To answer these questions, we employ a finite mixture model, and assume workers are earnings-maximizers. The findings show that informal wage employment has two distinct divisions each with a different wage-setting mechanism. Moreover, this results is robust to two definitions of informal wage employment – non-payment of employment income taxes, and absence of social security coverage. However, division size was found to be sensitive to how informal wage employment is defined. More specifically, the higher-paid and lower-paid divisions, by the social security criterion of informality, account for 86% and 14% of informal wage employment, respectively. But the composition changes to 72% higher-paid and 28% lower-paid if the non-payment of employment income tax definition is used. This major change implies that the two definitions suggest significantly different disaggregation of formal and informal wage employment. In addition, we found evidence in support of labour market segmentation, and estimated that up to 79% of all informal workers (found in both sections) would be better off in the formal sector or the other division of informal wage employment. Therefore, even though not all informal wage employment is inferior to formal work, for a majority of workers in that sector it is a sector of last resort. In this regard, expanding job opportunities in the formal sector would go a long way in reducing the incidence of informality.

The second essay answered two related questions; what determines the uptake of voluntary health insurance and to what extent does it affect health care utilization. The results indicate that older people, individuals from richer households and relatively better educated persons are more likely to voluntarily buy health insurance. This implies that poor individuals may be excluded from health insurance due, very likely, to poor information and their inability to afford premiums. Additionally, we use propensity score matching to examine the health care use (proxied by annual hospital visits) of individuals who buy voluntary health insurance with identical formal sector workers. The results indicate that voluntary health insurance is associated with increased health care use. This result must, however, be interpreted with caution because increased hospital visits does not necessarily imply higher health care costs. It is entirely possible that formal sector workers may seek more expensive care when they visit the hospital. Due to the absence of data on type of care sought and the associated costs we were unable to explore this hypothesis further. Future research can help us examine this question.

Finally, the last essay explores the magnitude of the gender wage gaps in the formal and informal sectors using quantile regression with a correction for selection using an identification-at-infinity strategy. The results indicate that the gender wage gap in the informal sector increases along the conditional wage function, but in the formal sector the wage gap is positive and larger at lower quantiles but disappears at the top. This implies that, in the formal sector, women earn higher wages than men in the lower end of the wage distribution, but this gap disappears at the top of the wage distribution. The decomposition result shows that up to 94% of the wage gap in the informal sector cannot be explained by human capital factors, but in the formal sector almost the entire wage gap can be explained by human capital factors. The higher than expected proportion of the unexplained wage gap in the informal sector partly hints at other institutional wage setting mechanisms in the sector which future research can help us better explain.

Bibliography

Acharya, A., S. Vellakkal, F. Taylor, E. Masset, A. Satija, M. Burke, and S. Ebrahim (2012). The impact of health insurance schemes for the informal sector in low-and middle-income countries: a systematic review. *World Bank Research Observer 28*(2), 236–266.

Addai, I. (2011). An Empirical Analysis of Gender Earnings Gap in the Ghanaian Informal Sector Using the 1998/1999 Ghana Living Standards Survey. *Pakistan Journal of Social Sciences 8*(3), 152–158.

Agesa, R. U. (1999). The Urban Gender Wage Gap in an African Country: Findings from Kenya. *Canadian Journal of Development Studies/Revue canadienne d'études du développement 20*(1), 59–76.

Aggarwal, A. (2010). Impact evaluation of India's 'Yeshasvini' community-based health insurance programme. *Health Economics 19*(S1), 5–35.

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics 84*, 488–500.

Albrecht, J., A. Björklund, and S. Vroman (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics 21*(1), 145–177.

Appleton, S., J. Hoddinott, and P. Krishnan (1999). The gender wage gap in three African countries. *Economic Development and Cultural Change 47*(2), 289–312.

Armitage, J. and R. Sabot (1991). Discrimination in East African's Urban Labor Market. *Unfair Advantage: Labor Market Discrimination in Developing Countries*, 75–94.

Arrow, K. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review 53*(5), 941–973.

Arrow, K. (1973). *The theory of discrimination*. Princeton University, Princeton.

Arrow, K. (1998). What has economics to say about racial discrimination? *Economic Perspectives 12*(2), 91–100.

Aryeetey, G. C., C. Jehu-Appiah, A. M. Kotoh, E. Spaan, D. K. Arhinful, R. Baltussen, S. van der Geest, and I. A. Agyepong (2013). Community concepts of poverty: an application to premium exemptions in Ghana's National Health Insurance Scheme. *Globalization and Health 9*(1), 12.

Asenso-Okyere, W. K., I. Osei-Akoto, A. Anum, and E. N. Appiah (1997). Willingness to pay for health insurance in a developing economy. a pilot study of the informal sector of Ghana using contingent valuation. *Health Policy 42*(3), 223–237.

Ashraf, J. and B. Ashraf (1993). Estimating the gender wage gap in Rawalpindi City. *Journal of Development Studies 29*(2), 365–376.

Axelson, H., S. Bales, P. D. Minh, B. Ekman, and U.-G. Gerdtham (2009). Health financing for the poor produces promising short-term effects on utilization and out-of-pocket expenditure: evidence from Vietnam. *International Journal for Equity in Health 8*(1), 20.

Baeza, C., F. Montenegro, and M. Nunez (2002). *Extending social protection in health through community based health organizations: evidence and challenges*. ILO, Geneva.

Battalio, R. C., J. H. Kagel, and K. Jiranyakul (1990). Testing between alternative models of choice under uncertainty: Some initial results. *Journal of Risk and Uncertainty 3*(1), 25–50.

Bauhoff, S., D. R. Hotchkiss, and O. Smith (2011). The impact of medical insurance for the poor in Georgia: a regression discontinuity approach. *Health Economics 20*(11), 1362–1378.

Becker, G. (1957). *economics of discrimination*. University of Chicago Press, Chicago.

Becker, G. (1975). Front matter, human capital: a theoretical and empirical analysis, with special reference to education. In *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, Second Edition*, pp. 22–0. NBER.

Becker, S. and A. Ichino (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal 2*(4), 358–377.

Bennett, S., A. Creese, and R. Monasch (1998). *Health insurance schemes for people outside formal sector employment*. WHO, Geneva.

Blau, F. D. and L. M. Kahn (2003). Understanding international differences in the gender pay gap. *Journal of Labor economics 21*(1), 106–144.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources 8*, 436–455.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of Applied Econometrics 13*(1), 1–30.

Cain, G. G. (1986). The economic analysis of labor market discrimination: A survey. *Handbook of Labor Economics 1*, 693–785.

Caliendo, M. (2006). *Microeconometric evaluation of labour market policies*, Volume 568. Springer Science & Business Media, Berlin.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge University Press, Cambridge.

Case, A. and A. Deaton (2003). *Consumption, health, gender, and poverty*, Volume 3020. Princeton University: Research Program in Development Studies.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics 32*(2), 189–218.

Chankova, S., S. Sulzbach, and F. Diop (2008). Impact of mutual health organizations: evidence from West Africa. *Health Policy and Planning 23*(4), 264–276.

Chen, W., X. Ying, S. Hu, L. Luo, and G. Sun (2002). Demand for supplementary health insurance. *Chinese Health Economics 21*(12), 17–21.

Chzhen, Y. and K. Mumford (2011). Gender gaps across the earnings distribution for full-time employees in Britain: Allowing for sample selection. *Labour Economics 18*(6), 837–844.

Coate, S. and G. C. Loury (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, 1220–1240.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics 24*, 295–313.

Cox, D. R. (1958). *Planning of experiments*. Wiley, New York.

Cremer, H. and P. Pestieau (1996). Redistributive taxation and social insurance. *International Tax and Public Finance 3*(3), 281–295.

Cutler, D. M. and R. J. Zeckhauser (2000). The anatomy of health insurance. *Handbook of Health Economics 1*, 563–643.

De Allegri, M., B. Kouyaté, H. Becher, A. Gbangou, S. Pokhrel, M. Sanon, and R. Sauerborn (2006). Understanding enrolment in community health insurance in sub-Saharan Africa: a population-based case-control study in rural Burkina Faso. *Bulletin of the World Health Organization 84*(11), 852–858.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics 84*(1), 151–161.

Dickens, W. T. and K. Lang (1985). A test of dual labor market theory. *American Economic Review 75*(4), 792–805.

Dimova, R., C. J. Nordman, and F. Roubaud (2010). Allocation of labor in urban West Africa: Insights from the pattern of labor supply and skill premiums. *Review of Development Economics 14*(1), 74–92.

DiPrete, T. A. and M. Gangl (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology 34*(1), 271–310.

Dixon, J., E. Y. Tenkorang, and I. Luginaah (2011). Ghana's National Health Insurance Scheme: helping the poor or leaving them behind? *Environment and Planning C: Government and Policy 29*(6), 1102–1115.

Dixon, J., E. Y. Tenkorang, I. N. Luginaah, V. Z. Kuuire, and G. O. Boateng (2014). National health insurance scheme enrolment and antenatal care among women in Ghana: is there any relationship? *Tropical Medicine & International Health 19*(1), 98–106.

Doeringer, P. B. and M. J. Piore (1985). *Internal labor markets and manpower analysis*. ME Sharpe, New York.

Dong, H., B. Kouyate, R. Snow, F. Mugisha, and R. Sauerborn (2003). Gender's effect on willingness-to-pay for community-based insurance in Burkina Faso. *Health Policy 64*(2), 153–162.

Ekman, B. (2004). Community-based health insurance in low-income countries: a systematic review of the evidence. *Health Policy and Planning 19*(5), 249–270.

Fan, C. S. and H.-K. Lui (2003). Structural change and the narrowing gender gap in wages: theory and evidence from Hong Kong. *Labour Economics 10*(5), 609–626.

Fields, G. (1975). Rural-urban migration, urban unemployment and underemployment, and job-search activity in LDCs. *Journal of Development Economics 2*(2), 165–187.

Fields, G. (1990). Labour market modelling and the urban informal sector: Theory and evidence.

Fields, G. (2004). *A guide to multisector labor market models*. Cornell University, New York.

Filmer, D. and L. H. Pritchett (2001). Estimating wealth effects without expenditure data- or tears: an application to educational enrollments in states of India. *Demography 38*(1), 115–132.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Gindling, T. H. (1991). Labor market segmentation and the determination of wages in the public, private-formal, and informal sectors in San Jose, Costa Rica. *Economic Development and Cultural Change 39*(3), 585–605.

Glewwe, P. (1996). The relevance of standard estimates of rates of return to schooling for education policy: A critical assessment. *Journal of Development economics 51*(2), 267–290.

Glick, P. and D. E. Sahn (1997). Gender and education impacts on employment and earnings in West Africa: Evidence from Guinea. *Economic Development and Cultural Change 45*(4), 793–823.

Gnawali, D. P., S. Pokhrel, A. Sié, M. Sanon, M. De Allegri, A. Souares, H. Dong, and R. Sauerborn (2009). The effect of community-based health insurance on the utilization of modern health care services: evidence from Burkina Faso. *Health Policy 90*(2), 214–222.

Gu, X. S. and P. R. Rosenbaum (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics 2*(4), 405–420.

Gumber, A. and V. Kulkarni (2000). Health insurance for informal sector: case study of Gujarat. *Economic and Political Weekly 30 September*, 3607–3613.

Günther, I. and A. Launov (2012). Informal employment in developing countries: Opportunity or last resort? *Journal of Development Economics 97*(1), 88–98.

Guo, S. and M. W. Fraser (2014). *Propensity score analysis: Statistical methods and applications*, Volume 11. Sage Publications, Los Angeles.

Harati, R. (2013). Heterogeneity in the Egyptian informal labour market: choice or obligation? *Revue d'Économie Politique 123*(4), 623–639.

Harris, J. R. and M. P. Todaro (1970). Migration, unemployment and development: a two-sector analysis. *American Economic Review 60*(1), 126–142.

Hart, K. (1973). Informal income opportunities and urban employment in Ghana. *Journal of Modern African Studies 11*(01), 61–89.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1), 153–61.

Heckman, J. (1990). Varieties of selection bias. *American Economic Review 80*(2), 313–318.

Heckman, J. and V. J. Hotz (1986). An investigation of the labor market earnings of Panamanian males evaluating the sources of inequality. *Journal of Human Resources 21*, 507–542.

Heckman, J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies 65*(2), 261–294.

Heckman, J., R. J. LaLonde, and J. A. Smith (1999). The economics and econometrics of active labor market programs. *Handbook of Labor Economics 3*, 1865–2097.

Heckman, J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics 30*(1), 239–267.

Hinks, T. (2002). Gender wage differentials and discrimination in the New South Africa. *Applied Economics 34*(16), 2043–2052.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Horton, S. (2002). *Women and industrialization in Asia*. Routledge, New York.

Huber, M. and B. Melly (2011). *Quantile regression in the presence of sample selection*. School of Economics and Political Science, Department of Economics, University of St. Gallen.

Hussmanns, R. (2004). Defining and measuring informal employment. *Bureau of Statistics Paper, ILO, Geneva*.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted SLS estimation of single-index models. *Journal of Econometrics 58*(1-2), 71–120.

ILO (1972). Employment, Income and Equality: A Strategy for Increasing Productivity in Kenya. Technical report.

ILO (1993). Statistics of employment in the informal sector, Report for the XVth International Conference of Labour Statisticians - Geneva. *Geneva: ILO Department of Statistics*.

ILO (2002). Women and Men in the Informal Economy: A Statistical Picture. Technical report.

ILO (2008). Global Employment Trends. Economic and Labour Market Analysis. Technical report.

ILO (2013). Statistical Update on Employment in the Informal Economy. Technical report.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*(1), 4–29.

Jehu-Appiah, C., G. Aryeetey, E. Spaan, T. De Hoop, I. Agyepong, and R. Baltussen (2011). Equity aspects of the National Health Insurance Scheme in Ghana: Who is enrolling, who is not and why? *Social Science & Medicine 72*(2), 157–165.

Jowett, M., A. Deolalikar, and P. Martinsson (2004). Health insurance and treatment seeking behaviour: evidence from a low-income country. *Health Economics 13*(9), 845–857.

Jowett, M. and W. C. Hsiao (2007). The Philippines: extending coverage beyond the formal sector. *Social Health Insurance for Developing Nations 434*, 81.

Jütting, J. P. (2004). Do community-based health insurance schemes improve poor people's access to health care? evidence from rural Senegal. *World Development 32*(2), 273–288.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Economic Perspectives 5*(1), 193–206.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*, 263–291.

Kelsey, D. and J. Quiggin (1992). Theories of choice under ignorance and uncertainty. *Journal of Economic Surveys 6*(2), 133–153.

King, G., E. Gakidou, N. Ravishankar, R. T. Moore, J. Lakin, M. Vargas, M. M. Téllez-Rojo, J. E. Hernández Ávila, M. H. Ávila, and H. H. Llamas (2007). A "politically robust" experimental design for public policy evaluation, with application to the mexican universal health insurance program. *Journal of Policy Analysis and Management 26*(3), 479–506.

Koenker, R. and G. J. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Kőszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics 121*(4), 1133–1165.

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research 33*(2), 188–229.

Lei, X. and W. Lin (2009). The new cooperative medical scheme in rural China: Does more coverage mean more service and better health? *Health Economics 18*(S2).

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *Manchester School of Economics and Social Studies 22*(2), 139–191.

Lieberman, S. S. and A. Wagstaff (2009). *Health financing and delivery in Vietnam: looking forward*. World Bank Publications.

Lindqvist, E. and R. Vestman (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *Applied Economics 3*(1), 101–128.

Loayza, N. and J. Rigolini (2006). Informality trends and cycles. *World Bank Policy Research Working Paper No. 4078*.

Lofgren, C., N. X. Thanh, N. T. Chuc, A. Emmelin, and L. Lindholm (2008). People's willingness to pay for health insurance in rural Vietnam. *Cost Effectiveness and Resource Allocation 6*(1), 16.

Lundberg, S. J. and R. Startz (1983). Private discrimination and social intervention in competitive labor market. *American Economic Review 73*(3), 340–347.

Machado, J. A. and J. Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics 20*(4), 445–465.

Magnac, T. (1991). Segmented or competitive labor markets. *Econometrica 59*, 165–187.

Maloney, W. F. (2004). Informality revisited. *World Development 32*(7), 1159–1178.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review 77*(3), 251–277.

Mas-Colell, A., M. D. Whinston, J. R. Green, et al. (1995). *Microeconomic theory*, Volume 1. Oxford University Press, New York.

Melly, B. (2005). Decomposition of differences in distribution using quantile regression. *Labour Economics 12*(4), 577–590.

Mensah, J., J. R. Oppong, and C. M. Schmidt (2010). Ghana's National Health Insurance Scheme in the context of the health MDGs: An empirical evaluation using propensity score matching. *Health Economics 19*(S1), 95–106.

Miller, G., D. M. Pinto, and M. Vera-Hernández (2009). *High-powered incentives in developing country health insurance: evidence from Colombia's Régimen Subsidiado*. National Bureau of Economic Research, Cambridge.

Mincer, J. A. (1974). The human capital earnings function. In *Schooling, Experience, and Earnings*, pp. 83–96. NBER.

Montenegro, C. (2001). *Wage distribution in Chile: Does gender matter? A quantile regression approach*. World Bank, Washington.

Nguyen, H. and J. Knowles (2010). Demand for voluntary health insurance in developing countries: the case of Vietnam's school-age children and adolescent student health insurance program. *Social Science & Medicine 71*(12), 2074–2082.

Nicodemo, C. (2009). Gender pay gap and quantile regression in European families. Technical report, IZA discussion papers.

Nketiah-Amponsah, E. (2009). Demand for health insurance among women in Ghana: cross-sectional evidence. *International Research Journal of Finance and Economics 33*.

Nordman, C. J. and F.-C. Wolff (2009). Is there a glass ceiling in Morocco? evidence from matched worker–firm data. *Journal of African Economies 18*(4).

Nyman, J. A. (1999). The value of health insurance: the access motive. *Health Economics 18*(2), 141–152.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review 14*(3), 693–709.

Osei-Boateng, C. and E. Ampratwum (2011). *The informal sector in Ghana*. Friedrich Ebert Stiftung, Accra.

Paulson, A. L. and R. M. Townsend (2005). Financial constraints and entrepreneurship: evidence from the Thai financial crisis. *Economic Perspectives 3Q*, 34–49.

Perry, G., W. Maloney, O. Arias, P. Fajnzylber, A. Mason, and J. Saavedra (2007). Informality: Exit and Exclusion, World Bank Latin America and Caribbean Studies. *World Bank, Washington DC*.

Peters, D. H. (2002). *Better health systems for India's poor: findings, analysis, and options*. World Bank Publications.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–661.

Polonsky, J., D. Balabanova, B. McPake, T. Poletti, S. Vyas, O. Ghazaryan, and M. K. Yanni (2009). Equity in community health insurance schemes: evidence and lessons from Armenia. *Health Policy and Planning 24*(3), 209–216.

Pratap, S. and E. Quintin (2006). Are labor markets segmented in developing countries? a semiparametric approach. *European Economic Review 50*(7), 1817–1841.

Preker, A. S., J. Langenbrunner, and M. Jakab (2002). Rich-poor differences in health care financing. *Social Reinsurance*, 21.

Psacharopoulos, G. and Z. Tzannatos (1991). *Women's Employment and Pay in Latin America*. World Bank, Washington.

Radchenko, N. (2014). Heterogeneity in informal salaried employment: Evidence from the Egyptian labor market survey. *World Development 62*, 169–188.

Rochet, J.-C. (1991). Incentives, redistribution and social insurance. *Geneva Papers on Risk and Insurance Theory 16*(2), 143–165.

Rosenbaum, P. R. (2002). Observational studies. In *Observational Studies*, pp. 1–17. Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rothschild, M. and J. Stiglitz (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics 90*, 629–649.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Saavedra, J. and A. Chong (1999). Structural reform, institutions and earnings: evidence from the formal and informal sectors in urban peru. *The Journal of Development Studies 35*(4), 95–116.

Salem, M. B. and I. Bensidoun (2012). The heterogeneity of informal employment and segmentation in the Turkish labour market. *Journal of the Asia Pacific Economy 17*(4), 578–592.

Samuelson, W. and R. Zeckhauser (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty 1*(1), 7–59.

Schippers, J. (1987). *Pay gap between men and women: economic analysis*. Wolters-Noordhoff [et al.], Groningen.

Schneider, P. and F. Diop (2001). *Synopsis of results on the impact of community-based health insurance on financial accessibility to health care in Rwanda*. World Bank Publications.

Schneider, P. and K. Hanson (2006). Horizontal equity in utilisation of care and fairness of health financing: a comparison of micro-health insurance and user fees in Rwanda. *Health Economics 15*(1), 19–31.

Schoemaker, P. J. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature 20*(2), 529–563.

Singh, K., I. Osei-Akoto, F. Otchere, S. Sodzi-Tettey, C. Barrington, C. Huang, C. Fordham, and I. Speizer (2015). Ghana's National Health insurance scheme and maternal and child health: a mixed methods study. *BMC Health Services Research 15*(1), 108.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics 87*(3), 355–374.

Stiglitz, J. E. (1974). Alternative theories of wage determination and unemployment in LDC's: The labor turnover model. *Quarterly Journal of Economics 88*(2), 194–227.

Stiglitz, J. E. (1976). The efficiency wage hypothesis, surplus labour, and the distribution of income in LDCs. *Oxford Economic Papers 28*(2), 185–207.

Tannuri-Pianto, M. and D. Pianto (2002). Informal employment in Brazil-a choice at the top and segmentation at the bottom: a quantile regression approach. *Anais do XXIV Encontro Brasileiro de Econometria 2*.

Thaler, R. H. and C. R. Sunstein (2003). Libertarian paternalism. *American Economic Review 93*(2), 175–179.

Townsend, R. M. (1994). Risk and insurance in village India. *Econometrica 64*, 539–591.

Trujillo, A. J., J. E. Portillo, and J. A. Vernon (2005). The impact of subsidized health insurance for the poor: evaluating the Colombian experience using propensity score matching. *International Journal of Health Care Finance and Economics 5*(3), 211–239.

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty 5*(4), 297–323.

Van de Ven, W. P. and B. M. Van Praag (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics 17*(2), 229–252.

Van Doorslaer, E., O. O'Donnell, R. P. Rannan-Eliya, A. Somanathan, S. R. Adhikari, C. C. Garg, D. Harbianto, A. N. Herrin, M. N. Huq, S. Ibragimova, et al. (2006). Effect of payments for health care on poverty estimates in 11 countries in asia: an analysis of household survey data. *The Lancet 368*(9544), 1357–1364.

Varian, H. R. (1992). *Microeconomic analysis*. Norton & Company, New York.

Verner, D. (1999). *Wage and productivity gaps: Evidence from Ghana*, Volume 2168. World Bank Publications.

Von Neumann, J. and O. Morgenstern (1944). *Theory of games and economic behavior*. Princeton University Press, Princeton.

Wagstaff, A. (2000). *Research on Equity, Poverty and Health: Lessons for the Developing World*. World Bank, Washington.

Wagstaff, A. (2007). *Health Insurance for the Poor: Initial Impacts of Vietnam's Health Care Fund for the Poor. Impact Evaluation Series# 11*. World Bank Publications.

Wagstaff, A. (2010). Estimating health insurance impacts under unobserved heterogeneity: the case of vietnam's health care fund for the poor. *Health Economics 19*(2), 189–208.

Weichselbaumer, D. and R. Winter-Ebmer (2005). A meta-analysis of the international gender wage gap. *Journal of Economic Surveys 19*(3), 479–511.

WorldBank (2000). World Development Report 2000-2001: Attacking Poverty. *World Bank*.

WorldBank (2004). *World Development Report 2005: Investment Climate, Growth and Poverty*. World Bank Publications.

Xu, K., D. B. Evans, K. Kawabata, R. Zeramdini, J. Klavus, and C. J. Murray (2003). Household catastrophic health expenditure: a multicountry analysis. *The Lancet 362*(9378), 111–117.

Yamada, G. (1996). Urban informal employment and self-employment in developing countries: theory and evidence. *Economic Development and Cultural Change 44*(2), 289–314.

Ying, X.-H., T.-W. Hu, J. Ren, W. Chen, K. Xu, and J.-H. Huang (2007). Demand for private health insurance in Chinese urban areas. *Health Economics 16*(10), 1041–1050.

Zhang, L., H. Wang, L. Wang, and W. Hsiao (2006). Social capital and farmer's willingness-to-join a newly established community-based health insurance in rural China. *Health Policy 76*(2), 233–242.

APPENDIX

APPENDIX TO CHAPTER 1

Table A.1: Information criteria for model selection

Informal sector	AIC	BIC
Homogeneous	40697.48	40780.30
2-division	30351.33	30528.87
3-division	30328.48	30597.66

Table A.2: Mixture model estimation with 2-division informal wage employment

	Formal	Informal-1	Informal-2
Constant	4.322 (36.55)***	3.994 (22.52)***	4.339 (30.48)***
Education *(ref: No school)*			
Low (1-6 years)	0.180 (1.41)	-0.156 (-1.05)	0.209 (1.59)
Medium (7-10 years)	0.375 (3.66)***	0.119 (0.91)	0.295 (2.66)*
High (> 10 years)	1.001 (10.03)***	0.236 (1.64)	0.652 (5.53)***
Age *(ref: 15–24)*			
25–44	0.585 (8.99)***	0.155 (1.92)***	0.340 (4.50)***
45 $\geq$	0.798 (11.01)***	0.084 (0.73)	0.445 (4.41)***
Other			
Location (Urban=1)	0.165 (3.80)***	0.315 (4.24)***	0.180 (2.80)*
Sex (Male=1)	0.190 (4.71)***	0.526 (7.12)***	0.551 (8.17)***
Marital status (Married=1)	0.265 (6.51)***	0.133 (1.98)**	0.251 (4.15)***
Industry			
Mining	0.642 (5.34)***	1.475 (10.57)***	0.337 (2.73)*
Manufacturing	-0.130 (-1.87)	0.259 (2.18)**	-0.223 (-2.07)**
Construction	0.096 (1.07)	1.020 (9.19)***	-0.097 (-0.85)

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

. . . continued

Table A.2: Mixture model estimation with 2-division informal wage employment, *continued ...*

	Formal	Informal-1	Informal-2
Commerce	-0.262 (-4.23)***	0.243 (2.63)*	-0.263 (-3.11)*
Transportaion & Storage & Communication	0.064 (0.97)	0.496 (4.15)***	-0.169 (-1.71)
π	0.493	0.144	0.363
Average earnings	808.51	242.20	453.08
N	2211	645	1624

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Table A.3: Distribution of individuals across divisions

	Formal		Informal-1		Informal-2	
	Value	[95% C.I]	Value	[95% C.I]	Value	[95% C.I]
π_j	0.493	[0.479, 0.508]	0.144	[0.134, 0.154]	0.363	[0.348, 0.377]
$\bar{\pi}_j$	0.778	[0.765, 0.790]	0.048	[0.042, 0.054]	0.174	[0.163, 0.185]
$\frac{\pi_j}{\bar{\pi}_j}$	0.635	[0.616, 0.654]	3.014	[2.588, 3.440]	2.079	[1.936, 2.223]

Table A.4: Distribution of workers across divisions where earnings will be maximized

	Formal		Informal-1		Informal-2	
Better-paid division	# of Workers	% Formal	# of Workers	% Informal	# of Workers	% Informal
Formal	1956	88.47	424	18.69	1104	48.66
Informal-1	37	1.67	71	3.12	106	4.67
Informal-2	217	9.86	150	6.61	414	18.25
N	2211	100	645	28.43	1624	71.57

Table A.5: Treatment effect (ATT)

Wage gaps	*NN* [a]	*Radius* [b]	*Kernel* [c]	*N*
Wage penalty	-487.9 ***	-336.0 ***	-435.9 **	1370
	(-3.79)	(-3.76)	(-2.22)	
Wage gain	226.8	141.6	266.6	410
	(1.34)	(0.80)	(1.45)	

Notes: Estimates obtained by bootstrap with 50 replications.

t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

[a] Number of neighbors used is 2.

[b] Caliper=0.001.

[c] Type of kernel is Gaussian, bandwidth (0.01)

Table A.6: T-test of covariate balance–Involuntary informal

	Mean		
	Treated	Control	Bias (%)
Education *(ref: No school)*			
Low (1-6 years)	0.053	0.045	3.6 (0.66)
Medium (7-10 years)	0.295	0.273	4.8 (0.90)
High (>10 years)	0.620	0.654	-6.9 (-1.28)
Age *(ref: 15–24)*			
25–44	0.637	0.676	-1.1 (-0.16)
45 $\geq$	0.231	0.218	3.1 (0.57)
Industry			
Mining	0.032	0.031	0.8 (0.15)
Manufacturing	0.080	0.084	-1.2 (-0.22)
Construction	0.053	0.045	3.7 (0.69)
Commerce	0.156	0.156	0.2 (0.03)
Transportaion & Storage & Communication	0.146	0.133	3.7 (0.68)
Other			
Location(Urban=1)	0.806	0.817	-2.8 (-0.53)
Sex(Male=1)	0.724	0.710	3.0 (0.56)
Marital status(Married=1)	0.663	0.672	-1.9 (-0.34)
Household head	0.743	0.756	1.6 (0.30)
Household size	3.691	3.624	3.1 (0.57)
Infants in HH	0.060	0.074	-5.7 (-1.05)
Children in HH	1.183	1.122	4.6 (0.86)
HH members aged $\geq$ 60	0.076	0.080	-1.5 (-0.27)
N	173	173	

Note Test based on kernel matching results
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Table A.7: T-test of covariate balance–Voluntary informal

	Mean		
	Treated	Control	Bias (%)
Education *(ref: No school)*			
Low (1-6 years)	0.093	0.102	-3.3 (-0.33)
Medium (7-10 years)	0.780	0.771	2.3 (0.24)
High (>10 years)	0.020	0.020	-0.0 (-0.00)
Age *(ref: 15–24)*			
25–44	0.624	0.615	2.0 (2.0)
45 $\geq$	0.039	0.039	-0.0 (-0.0)
Industry			
Mining	0.020	0.010	8.1 (0.82)
Manufacturing	0.102	0.102	-0.0 (-0.00)
Construction	0.185	0.185	-2.1 (-0.21)
Commerce	0.317	0.327	-0.7 (-0.07)
Transportaion & Storage & Communication	0.210	0.210	0.0 (-0.00)
Other			
Location (Urban=1)	0.649	0.649	0.0 (-0.0)
Sex (Male=1)	0.727	0.717	2.2 (0.22)
Marital status (Married=1)	0.249	0.259	-2.2 (-0.23)
N	424	424	

Note Test based on kernel matching results
t statistics in parentheses ** $p < 0.05$, * $p < 0.01$, *** $p < 0.001$

Figure A.1: Kernel density of monthly earnings

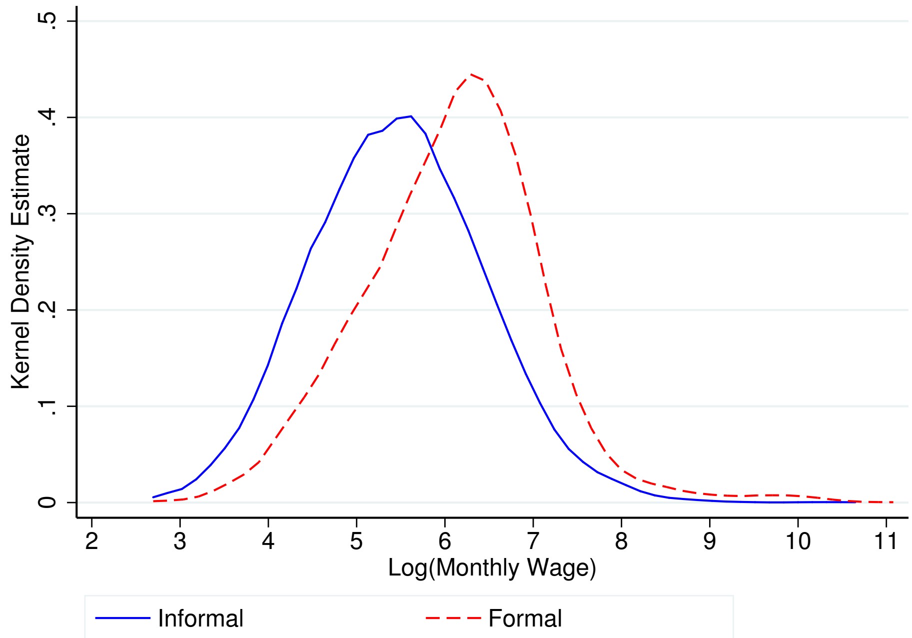


Figure A.2: Distribution of workers across divisions

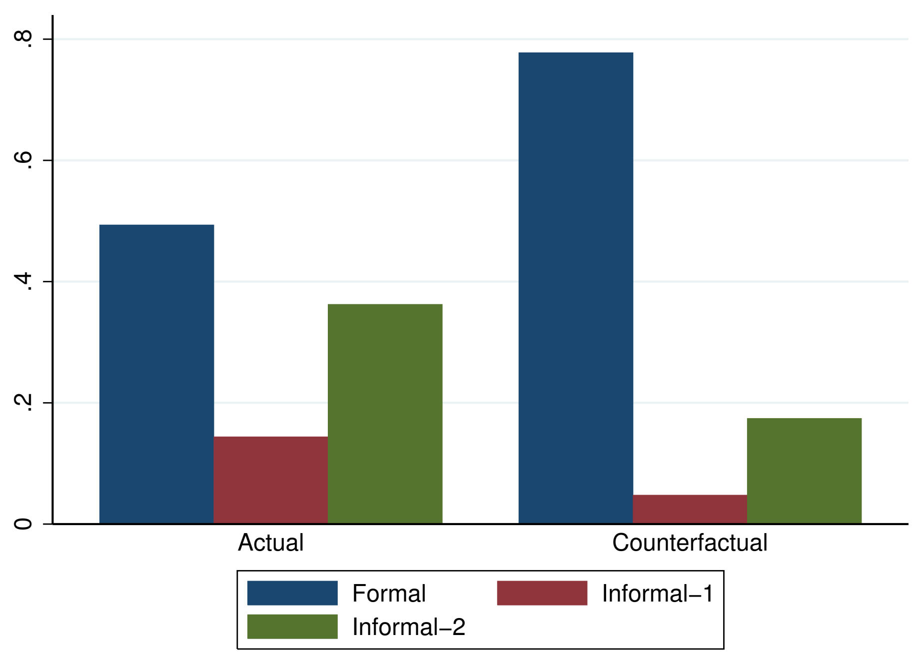


Figure A.3: Propensity score distribution before and after matching–Involuntary informal vs formal

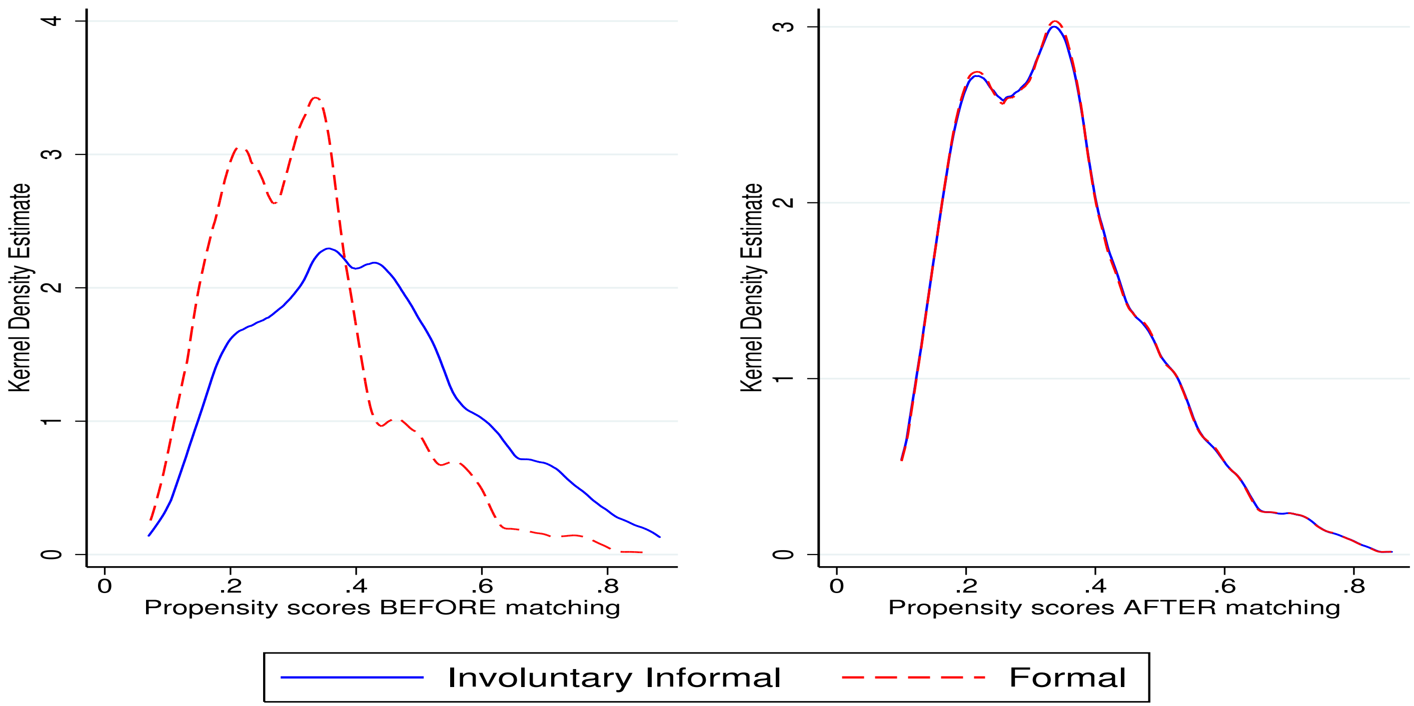


Figure A.4: Propensity score distribution before and after matching–Voluntary informal vs formal

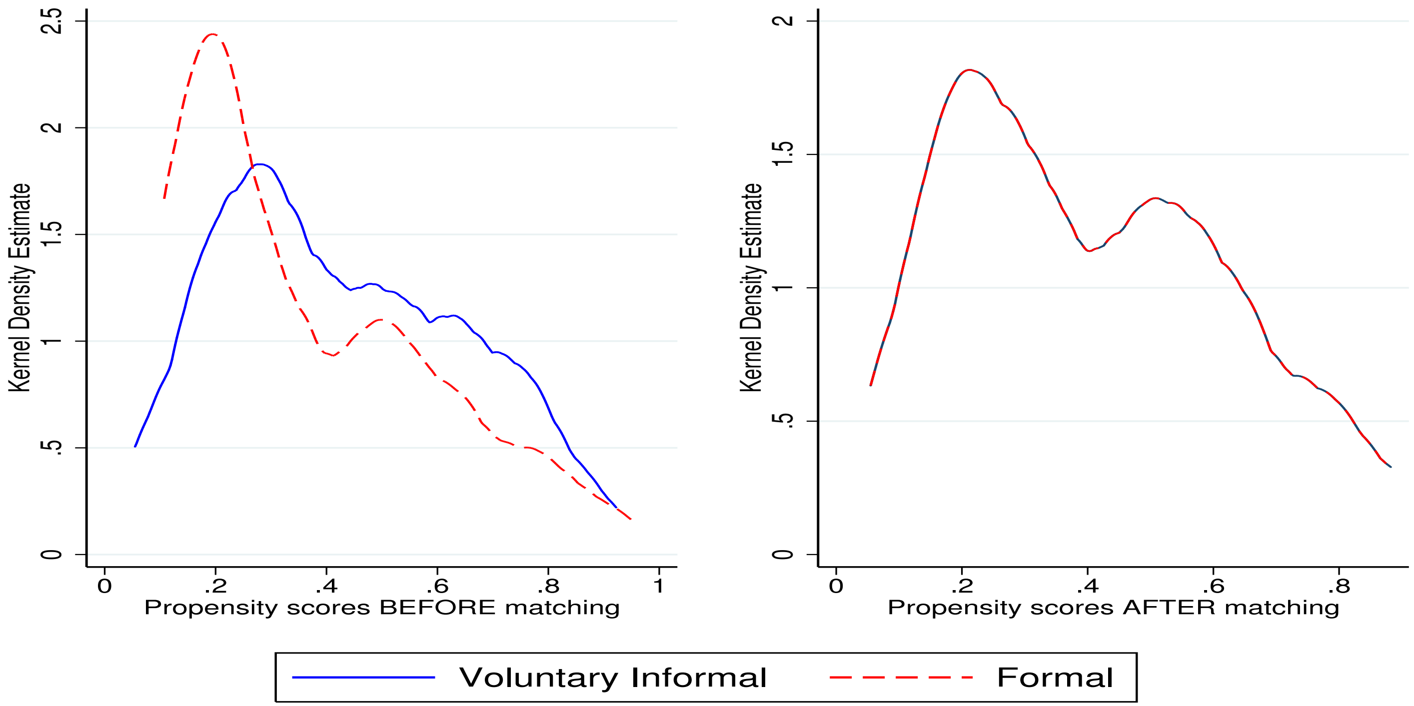


Figure A.5: Test of covariate balance–Involuntary informal vs formal

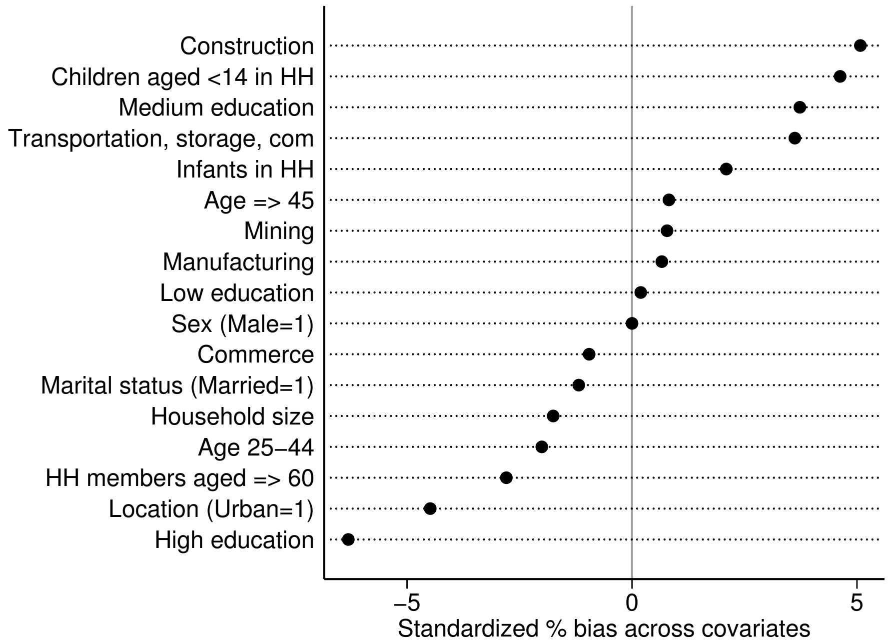


Figure A.6: Test of covariate balance–Voluntary informal vs formal